DESIGNING ROBUST MEASUREMENT NETWORKS USING UNIVERSAL MULTIPLE LINEAR REGRESSION

by

Melissa Clutter

_____

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF HYDROLOGY AND ATMOSPHERIC SCIENCES

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2019

2019

THE UNIVERSITY OF ARIZONA

GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Melissa Clutter titled *Designing Robust Measurement Networks using Universal Multiple Linear Regression* and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Dr. Paul "Ty" Ferre — Date: 5/20/19

Dr. Thomas Meixner — Date: 5-20-2019

Dr. Marcel Schaap — Date: 5-20-2019

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Dr. Paul "Ty" Ferre — Date: 5/20/19

Dissertation Committee Chair

Hydrology and Atmospheric Sciences

Acknowledgements

I have received a great deal of support throughout my dissertation process. I would first like to thank my primary research advisor Dr. Ty Ferre for teaching me how to be a better academic by helping me formulate an engaging research topic and always pushing me to explore more deeply.

I also would like to thank Dr. Hoshin Gupta, Dr. Thomas, Meixner, and Dr. Marcel Schaap for serving on my dissertation committee. Their contributions and feedback were invaluable. They intellectually challenged me while also encouraging me to think more deeply. Additionally, Jeff Klakovich was my MODFLOW and PEST "guru". The parameter estimation section of my dissertation (Paper 3) would not have been possible without his help.

Lastly, I would like to thank my friends and family for supporting me. I am fortunate to have several role models that demonstrate how to have a healthy work-life balance. Some of my friends encourage me to work on the weekends and some inspired me to play outside and rock climb when I was too stressed. I feel that my friends and family always knew I could finish my doctoral degree and have fun at the same time. Even when I didn't know if it was possible.

Table of Contents

Dissertation Abstract

Collecting hydrological data is essential for understanding system behavior and processes; without it, there is no basis for predictive modeling or risk assessment. Unfortunately, limited monitoring budgets often restrict measurement designs for field-based studies. Therefore, most field studies require some sort of data worth analysis to identify the most important data to collect with respect to the prediction(s) of interest. Data worth analyses can be either informal using methods such as trial-and-error, intuition, or rules of thumb, or formal using a quantitative metric to identify the most valuable data. My research focuses on a simple, computationally inexpensive formal data worth analysis which can be used in conjunction with more complex optimization approaches or when they are not warranted.

A key to network design is that the selection of sensor type, timing, and placement should be both informative and efficient. There are many possible individual sensor types and installation depths, and the key is to determine which sets of observations would be most effective prior to data collection. My research explores a combination of a method called universal Multiple Linear Regression (uMLR) and Robust Decision Making (RDM) to identify these best observation sets. The uMLR method quantifies the explanatory power of all possible combinations of observations to the prediction(s) of interest and the RDM strategy further explores the impacts of user-defined uncertainties, including measurement error and parameter uncertainty, on these observation-set selections. Robust Decision Making is a concept developed by the Research and Development (RAND) Corporation and is designed to select a robust outcome under a range of uncertainty, at the risk of the selection being sub-optimal for any one specific uncertain outcome. Norgaard et al., (2014) previously used the uMLR approach to downsampling pre-existing data to identify a reduced set of parameters to describe the dispersibility of colloids. I offer an extension of the uMLR downsampling approach, based on model-simulated data, to consider optimizing data that have not yet been collected.

Introduction

My research explores the combined uMLR and RDM method in three incremental stages and each stage is discussed in its own chapter of the dissertation. The first chapter introduces the predictive approach to uMLR combined with RDM. HYDRUS 1D is used to simulate virtual data and the outcomes of interest needed to perform uMLR, and an ensemble of models is used to explore known sources of uncertainty such as measurement error and uncertain structure/parameters. The paper recognizes uMLR combined with RDM as a computationally inexpensive method that could be used for quantifying the benefits of different observation sets under user-defined uncertainty. The paper was accepted for publication in the Soil Science Society of America Journal on February 1st, 2019.

The second chapter extends the combined uMLR and RDM method to field data from the Prototype Hanford Barrier (PHB) site. An existing data set of neutron probe measurements from the PHB site is downsampled using uMLR. Additionally, data simulated in HYDRUS 1D are used in a predictive approach using uMLR. The optimized observation depths for the downsampling and predictive approaches are compared. This analysis confirms the potential benefits of uMLR for network design prior to sensor installation. This paper also introduces the idea of selecting ranges of sensor placements for more practically applicable recommendations. The paper has been submitted for publication in Water Resources Research

The third chapter explores the possibility of optimizing data collection for parameter estimation. Data is simulated in MODFLOW and the uMLR method is used to find the best observations for inferring hydraulic conductivity of two layers in a multilayer system using the attenuation of atmospheric pressure signals. The suggested pressure observations from uMLR are used in a parameter estimation software (PEST) to estimate the hydraulic conductivity. This represents several advances of the uMLR method. First, multiple predictions of interest are considered, necessitating the use of multivariate regression. Second, uMLR is applied to identifying system parameters, rather than system states. Third, the relationship between the prediction(s) of interest and observations is nonlinear.

Chapter 1

Designing robust, cost-effective field measurement sets using universal multiple linear regression

Melissa Clutter and Ty Ferré

Clutter, M., and T. P.A. Ferré (2019). Designing Robust, Cost-Effective Field Measurement Sets using Universal Multiple Linear Regression. *Soil Science Society of America Journal.* 0nline. doi:10.2136/sssaj2018.09.0340.

Abstract

Limited monitoring budgets restrict the type and number of sensors that can be installed for field-based studies. Therefore, sensor selection should be both informative and efficient. We propose a method to optimize sensor network design, prior to data collection, by combining multiple linear regression and robust decision-making (RDM). Multiple linear regression inherently considers the strength of the relationship between observations and predictions of interest and correlations among proposed observations. In our approach, we use universal Multiple Linear Regression (uMLR) to quantify the explanatory power of all possible combinations of model-simulated candidate observations (of different sensor types and locations). A model-ensemble approach allows for network design in the context of user-defined uncertainties, including expected measurement error and parameter and structural uncertainty. Application of uMLR with RDM produces a comprehensive assessment of the likely value of many observation sets. These results can be used to design sensor networks to address specific experimental objectives and to balance the cost and effort of installing sensors to the expected value of the data for model testing and decision support.

Introduction

Most hydrogeologic investigations have limited budgets for data collection (James and Gorelick, 1994; Mogheir et al., 2003; Hubbart et al., 2005; Liu et al., 2012). Data collection designs typically rely on some form of data worth analysis to identify the most valuable data with respect to a prediction of interest. That is, given current understanding – including insights regarding known uncertainties – experimentalists aim to allocate sampling resources most efficiently. Data worth analyses include both informal methods such as trial-and-error, rules of thumb, or intuition, and formal methods using a quantitative metric to identify useful data (James and Gorelick, 1994; Strebelle, 2002; Mogheir et al., 2003; Tiedeman et al., 2003; Çamdevýren et al., 2005; Fernandez-Galvez et al., 2006; Vereecken et al, 2008; Jolliffe, 2011; Sun et al., 2013; Hermans et al., 2015; Kikuchi, 2017; Lin et al., 2017; Ju et al., 2018; Vilhelmsen and Ferre, 2018). To some degree, every experimental design relies on some mixture of these approaches. One goal of this study is to propose a formal method of measurement design that is simple and computationally inexpensive to implement, making it more likely to be used than more complete, but more complex data worth analyses. We expect that this will be particularly useful for relatively simple experiments that may not warrant advanced measurement optimization approaches.

In this study, we consider a relatively simple example problem: quantifying the total water stored in a depth interval in the vadose zone. Currently, no measurement method can provide a single measure of the total length of water in a subsurface interval. The most direct way to determine this value is to monitor the water content repeatedly with high vertical resolution and to sum these observations at each measurement time. This could be achieved with a neutron probe (Gardner and Kirkham, 1952).

However, restrictions related to the radioactive source may preclude the use of this instrument and, depending upon the desired temporal resolution, reoccupying the site at each measurement time may be impractical. Alternatively, high resolution profiling could be achieved with buried sensors, such as time-domain reflectometry probes (Topp et al., 1980). But, the cost of installing the large number of sensors needed to achieve high depth resolution may be prohibitive. This leads naturally to two related questions: how many point sensors are necessary to measure time-varying water storage in the vadose zone and at what depths should these sensors be installed to provide the most accurate measure of the total water stored?

The question of adequate spatial resolution of water content measurements to define the total water stored is relatively easy to define in retrospect - after data have been collected. A downsampling exercise can be performed to remove redundant observations (Reed et al., 2000a; 2000b). Furthermore, the weight that should be applied to each remaining sensor can be determined based on a multiple linear regression (MLR) between the remaining sensors and the outcome of interest, total length of water, as determined with the full data set. This method was developed by Norgaard et al. (2014) for a different measurement optimization context. Specifically, they used an approach that they referred to as universal multiple linear regression (uMLR) to downsample many *previously-collected* measurements to identify a reduced set of measurements that could have been used as a surrogate for more difficult or expensive analyses. Specifically, in developing uMLR, Norgaard et al. (2014) sought to examine the ability to use highly disparate soil analyses including soil electrical conductivity, water content at specific pressures, and soil particle size fractions as indicators for other, more costly analysis – colloid mobilization and transport. Norgaard et al. (2014) had no reason to expect that colloid dispersibility and transport would have a linear dependence on any of the other measurements that they collected. But, they recognized that a linear dependence is not necessary for design optimization. Rather, the existence of such a linear relation could allow for the inference of the target observation using other observations as surrogates. The advantage of uMLR is that it considers all possible combinations of the previously collected observations of any given subset size, which allows for flexibility in the measurement network design. For example, multiple observation sets could be assessed based on the effort required to collect the data or the cost of the analyses, allowing for cost-benefit optimization. Or, the ability of observation sets to allow for the inference of several measurement targets could be considered to fulfill multiple experimental objectives.

We extend the uMLR approach to consider data that have not yet been collected. That is, we show how uMLR can be used for predictive design in addition to downsampling previously collected data. Specifically, we show how a user can make objective choices about measurement network design in the context of user-defined uncertainties prior to sensor installation, thereby optimizing the type, number, and locations of sensors to measure an outcome of interest. In this mode, uMLR can be seen as a simple version of more comprehensive (and more computationally expensive) data worth analyses (e.g. James and Gorelick, 1994; Strebelle, 2002; Mogheir et al., 2003; Tiedeman et al., 2003; Çamdevýren et al., 2005; Fernandez-Galvez et al., 2006; Gupta et al., 2008; Vereecken et al, 2008; Jolliffe, 2011; Gupta et al., 2012; Sun et al., 2013; Hermans et al., 2015; Kikuchi, 2017; Lin et al., 2017; Ju et al., 2018; and Vilhelmsen and Ferre, 2018). There are three specific advantages of our predictive implementation of uMLR. First, the low computational effort and ease of use make uMLR accessible even for relatively simple experiments (which likely face the most severe budgetary limitations). Second, uMLR allows for experimentalists to specifically consider the impact of different sources of uncertainty on measurement

network design prior to sensor installation. Third, uMLR provides multiple measurement sets with associated metrics of quality, which allows for cost-benefit analyses as a part of network design.

Methodology

The uMLR method relies on proposing a simple linear relationship between all of the possible combinations of user-proposed candidate observations and the outcome(s) of interest. Like a simple trained neural network (Haykin, 1994), the uMLR approach determines whether a weighted linear combination of observations provides an adequate estimate of an outcome of interest. The goodness of fit of a linear regression (R^2) of each set of observations to the outcome of interest provides a measure of the adequacy of that set of measurements to infer the outcome of interest (Draper and Smith, 2014). If the outcome of interest cannot be approximated adequately from a linear combination of the observations, then the R^2 will be low, indicating that the set of observations is not acceptable. This metric can be used to rank observation sets and to determine if any proposed sets provide sufficient accuracy to satisfy the experimental objectives.

In this study, we extend uMLR so that it can be used for predictive design, prior to sensor installation. There are two insights that allow for this extension. First, forward models can be used to provide the virtual data and outcomes of interest needed to perform uMLR. Second, an ensemble of models, representing known sources of uncertainty, can be used as the basis of designing robust measurement networks under proscribed uncertainty.

Predictive design with uMLR involves five steps. First, the outcome(s) of interest and all candidate observations (type, depth, etc.) are defined by the user. Second, a forward model of the processes under study is developed. Third, all sources of known uncertainty are defined, including those related to measurement uncertainty, initial and boundary conditions, parameter values, and model structure. Fourth, an ensemble of models is developed to represent the processes and all known uncertainties and is used to simulate candidate observations and the corresponding outcome(s) of interest. Fifth, uMLR is applied to all observation sets (for all uncertain scenarios) to assess their performance over the entire model ensemble.

The first four steps of our application of uMLR are relatively simple extensions of the analysis presented by Norgaard et al. (2014), albeit applied to model-derived responses rather than previously-collected data. The main advance presented here is the optimization over the model ensemble. For this, we chose to use a quantitive decision analytic framework designed to evaluate strategies under uncertainity referred to as Robust Decision Making (RDM) (Lempert et al., 2013). Robust Decision Making is a concept developed by the Research and Development (RAND) Corporation. It was primarily designed to support decision-making by seeking a common design across many plausible future scenarios. The caveat is that the RDM-selected design is robust under uncertainty at the risk of being sub-optimal for any specific scenario. RDM has been used in high-level policy processes such as flood risk management decisions in the United Kingdom (Hine and Hall, 2010), long range water management plans for the western United States (Groves et al., 2008; Lempert and Groves, 2010) and for energy policies in Israel (Popper et al., 2009). We apply RDM by performing uMLR over each model in the ensemble to consider all known uncertainties. We determine the goodness of fit (R^2) for each observation set for each model. Then we identify the minimum R^2 for each set over the model ensemble. Consistent with the RDM approach, the observation set with the highest minimum R^2 over the ensemble provides the most robust design under uncertainty. If the R^2 of this design is acceptably high, then the uMLR analysis suggests that

it can be used for the monitoring objective, subject to the assumptions underlying the models used to represent the system and the known uncertainties considered.

The uMLR approach is computationally inexpensive; however, the nature of combinations can lead to an unacceptably large number of sets to consider. Specifically, the number of subsets, p_{set}, that can be formed from c_o candidate observations with a maximum set size of s_{max} is:

$$p_{set} = \sum_{n=1}^{s_{max}} \frac{c_o!}{n!*(c_o-n)!} \quad (1)$$

For example, considering all combinations of 20 candidate observations for observation sets ranging from 1 to 5 members totals to 21,699 sets. But, considering sets up to 10 members totals to 616,665 sets. Increasing the number of candidate observations to 50 for sets of up to 5, 10, and 30 members totals to 174,436, 5.3E07, and 1.07E09 sets, respectively. These calculations can be conducted in parallel, if necessary. But, in general, it is worth thinking carefully about the number of candidate observations and the maximum number of observations considered in a set before conducting the uMLR analysis.

Case study

In semi-arid regions, a significant amount of water is thought to be stored in the shallow subsurface beneath ephemeral streams. As a result, these features play an important role in supporting riparian habitats during interflow periods (Walvoord et al., 2004). For simplicity of illustration, we consider an idealized ephemeral stream that experiences flow for one hour every day in response to localized, heavy precipitation followed by 23 hours with no surface flow. The goal of the analysis was to illustrate how uMLR can provide a network design that is both efficient and sufficient for monitoring dynamic water storage in the upper 2 m of a channel beneath an ephemeral stream.

Candidate observations and the outcome of interest

The simplest and most obvious monitoring network to quantify water storage within a subsurface depth interval would be comprised of water content observations. We consider up to 20 possible water content sensor depths (from 0 to 190 cm with a 10 cm interval). This range corresponds with the depth (2 m) within which we sought to quantify the total water storage on an hourly basis. One advantage of the uMLR approach is that it allows for relatively simple and efficient examination of the potential value of multiple measurement types. For this study, we also consider temperature sensors. This proposed addition to the monitoring network is based on previously demonstrated use of temperature monitoring to infer water flux in the subsurface (Bredehoeft and Papadopulos, 1965; Stallman, 1965; Anderson, 2005; Keery et al., 2007). We consider 10 possible temperature sensor depths (from 0 to 90 cm with a 10 cm interval). The shallower range was determined by previous work that indicated shallower temperature measurements are preferred for similar infiltration conditions (Clutter and Ferré, 2018) because deeper observations are likely to lie below the extinction depth of temperature variation (Dickinson et al., 2014).

The outcome of interest for this analysis is the total length of water stored within a given depth range in the vadose zone as a function of time. We consider the case for which there is a limited budget for monitoring equipment. Specifically, up to five sensors could be installed (s_{max}: Equation 1), comprised of

any combination of 30 temperature or water content observations (c_o: Equation 1). It is assumed that data are collected hourly.

Based on Equation 1, all sets of up to 5 members, selected from the 30 candidate observations, results in 174,436 possible sets. The uMLR method requires approximately 0.005 seconds to assess each observation subset on a single processor, regardless of the number of observations in the set. Therefore, 174,436 possible sets only require approximately 15 minutes to analyze for each model in the ensemble. Note that this computational time is in addition to the time required to run the models used to produce the synthetic data. A more complex forward model could easily require more computational effort that the uMLR analysis. But, the number of observation sets considered during uMLR is independent of the forward model run time. In contrast, consideration of many sources of uncertainty, leading to a larger model ensemble, will cause a linear increase in the computational effort of both the forward model and the uMLR analysis.

System and model description

Using HYDRUS 1D (Šimunek et al., 2008) we model a stream channel that consists of homogenous soil and contains native grass and shrubs with uniformly distributed roots between 25-100 cm depth. Flow is modeled as 1D, vertical. The default van Genuchten (1980) soil hydraulic properties for clay, silt, silty loam, loam, and sandy loam were provided by Schaap et al. (2001); these properties were not considered to be temperature dependent. For water flow, the top boundary condition was a constant pressure head of 0 cm from 12:00 a.m. to 1:00 a.m. to represent zero-ponded height flooding, and zero flux for the other 23 hours of the day to represent no flow conditions. The bottom boundary condition was free drainage.

The potential evapotranspiration rate was 0.4 cm/day (Kurc and Small, 2004). Using a built-in HYDRUS capability, the ET was constant and equal to 1% of the total daily value between 0-6 a.m. and 6 p.m.-midnight and had a sinusoidal shape for the rest of the day, with a maximum at 12 noon. Root water uptake was described using the Feddes et al., (1976) model and the following parameters based on typical desert shrubs (Kurc and Small, 2004): anaeobiosis pressure head of -5.4 cm, optimal uptake between -22.4 cm and -32.6 cm, the water uptake decreases linearly from 0.3 to 0.05 cm/day between -32.6 cm and -800 cm, and uptake is zero for pressure heads less than the wilting point of -1557.4 cm.

The temperature at the top boundary varied sinusoidally between 5-35 °C with a 24-hour period and a maximum at 1 p.m. as represented by the sine function in Kirkham and Powers (1972). The bottom heat transport boundary condition was zero gradient. The initial temperature conditions were equal to the mean daily temperature of 20 °C. The default HYDRUS 1D thermal parameters were used based on Chung & Horton (1987).

To provide the simplest dynamic system for demonstration of uMLR, we considered oscillatory steady state conditions. To achieve this, oscillatory steady state was reached within 5-30 days, depending upon the soil type, until the water content at each depth was consistent at any given time of day (although varying over a daily cycle). For consistency, hourly observations from day 30 were used for all soils. The initial and minimum time step of the model was 0.0001 days. The water content and temperature were simulated at each proposed measurement depth. HYDRUS also reports the total length of water in the profile, providing the target for the uMLR analysis (outcome of interest). While we chose simple

transient conditions for ease of discussion, any other dynamic conditions could be considered following the same procedure.

To facilitate discussion of the effects of uncertainty on measurement network design, we consider increasing levels of uncertainty. First, to demonstrate a direct application of uMLR as developed by Norgaard et al. (2014) to the example problem, we show how a measurement network can be optimized in the absence of known uncertainty. Second, we apply uMLR with RDM for predictive design considering only measurement error. Finally, we apply uMLR with RDM considering measurement uncertainty and uncertainty in the soil hydraulic parameters.

Results and discussion

Examples of the simulated hourly water content (*Figure 1*) and temperature (*Figure 2*) at the candidate depths represent the data to be analyzed. The total storage (*Figure 3*a) is the combined result of three transient process: infiltration, evapotranspiration, and deep drainage, as shown in Figure 3b.

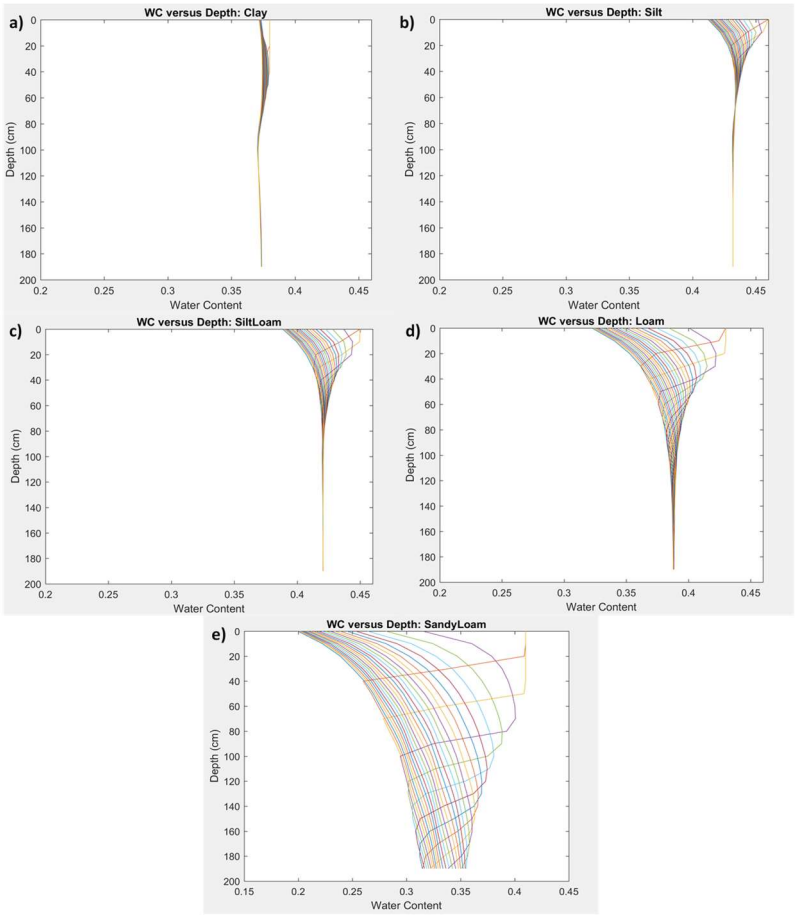

Figure 1: Data simulated using HYDRUS 1D. Water content vs. Depth for the 5 soil types. Each line represents 1 hour during dynamic steady-state for a total of 24 hours.

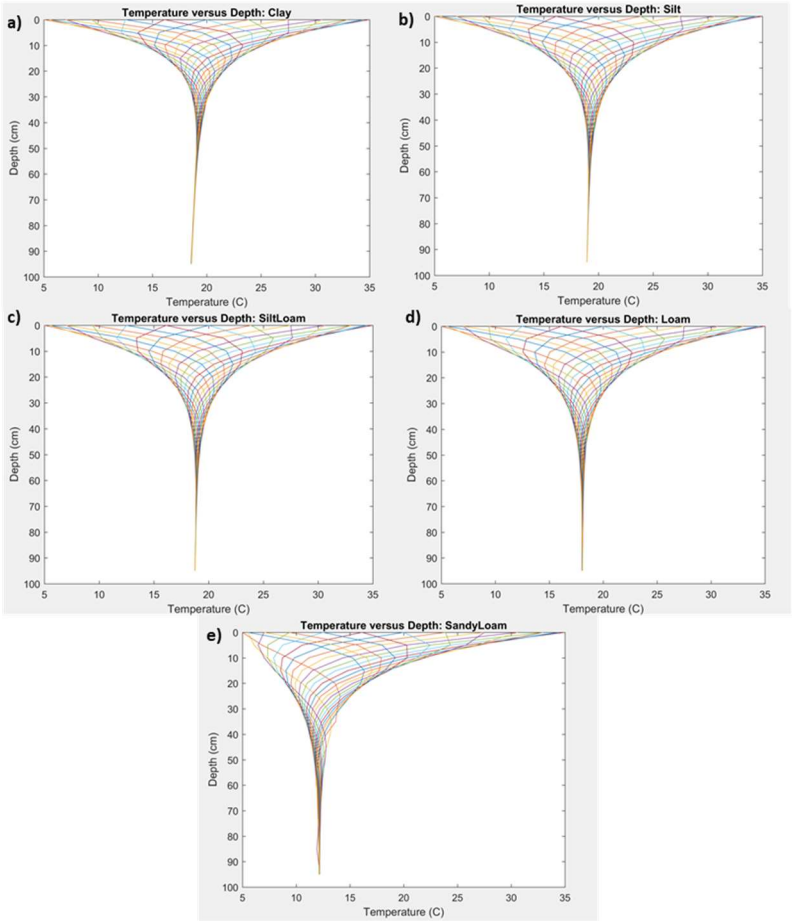

Figure 2: Data simulated using HYDRUS 1D. Temperature vs. Depth for the 5 soil types. Each line represents 1 hour during dynamic steady-state for a total of 24 hours.

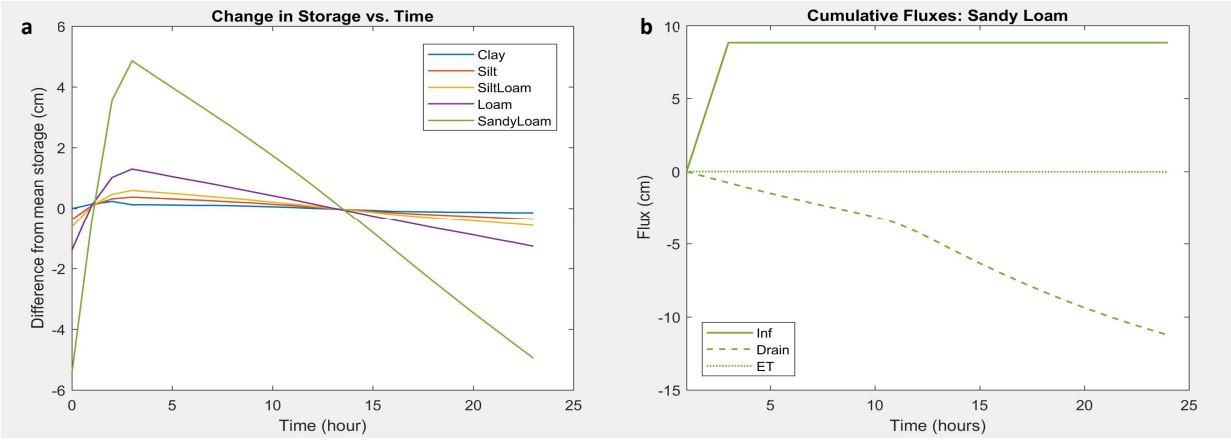

Figure 3: Data simulated using HYDRUS 1D. a) The change in total hourly storage (cm) versus time for each of the 5 soil types during dynamic steady-state. Infiltration happens during hour 1 and the infiltration flux is zero for the other 23 hours of the day. b) Example of the cumulative fluxes (infiltration, drainage, and ET) that are contributing to the change in storage for a sandy loam. Fluxes entering the domain are positive; fluxes leaving the domain are negative.

It is conceptually simple to imagine using water content measurements to define the total water stored in the vadose zone. But it is instructive to consider how a limited number of water content sensors should be distributed for optimal monitoring. For our hydrologic system, the greatest variation in water content occurs at the surface and dampens with depth, with most soils showing no change in water content below 160 cm (Figure 1). Intuitively, we would expect to have one or more shallow sensors to capture this variability. But, to reduce redundancy, these should not be placed too close to one another. As we add more sensors, we may also expect to include one or more deep sensors to define the non-varying water content at depth. (Note: a deep observation is necessary to define the total water stored, but it would not be necessary if only the temporal change in storage was of interest). After some number of sensors have been added, additional water content sensors will offer diminishing returns as they become increasingly redundant with other sensors. Finally, we can assume that the larger the measurement uncertainty, the more sensors will be needed to give an accurate estimation of the total length of water in the profile.

These general rules are simple to understand and lead to a 'typical' design of equally spaced sensors or sensors with a marginally smaller separation near the surface. But, it is difficult to make specific, quantitative recommendations about the number, types, and depths of sensors that should be used. For example, the greatest temperature variation with time occurs at shallow depths, with most soils showing no change below 70 cm depth (*Figure 2*). How should we combine this information with that used to intuitively select water content measurement depths, as discussed above? Using uMLR, we can compare the explanatory power of different proposed observation sets considering both measurement sensitivity (Gupta et al., 1998, 2008, 2012) and measurement redundancy in the specific context of the outcome of interest. Furthermore, we can extend the analysis to consider different measurement types (e.g water content and temperature).

Optimizing water content observations - sandy loam, no uncertainity

We first analyzed the relationship between water content observations and total water stored for the simplest, most well characterized conditions: the soil type was known; only water content observations were considered; and there was no measurement error (*Figure 4*a; Figure 5). We consider observation sets of up to 5 sensors out of 20 candidate water content sensor depths. According to Equation 1, this leads to 20 sets comprised of one sensor. Sets of two through five sensors have: 190; 1,140; 4,845; and 15,504 possible observation sets, respectively. The R^2 of each subset is plotted against the number of sensors in the set on *Figure 4*a. The goodness of fit and the specific sensors selected for the best set of sensors (highest R^2) for each observation set size are shown on Figures 5a and 5b, respectively.

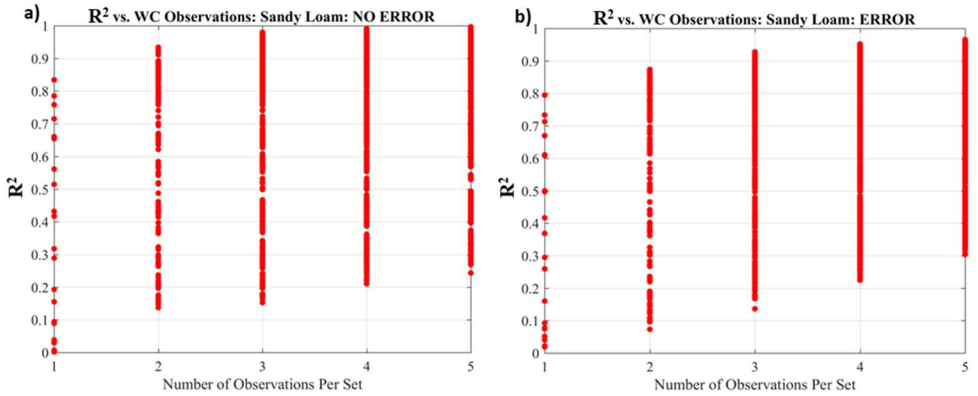

Figure 4: a) R^2 for best sets ≤ 5 without noise and b) the RDM R^2 with noise. Each point represents one subset of the possible observations when considering all combinations.

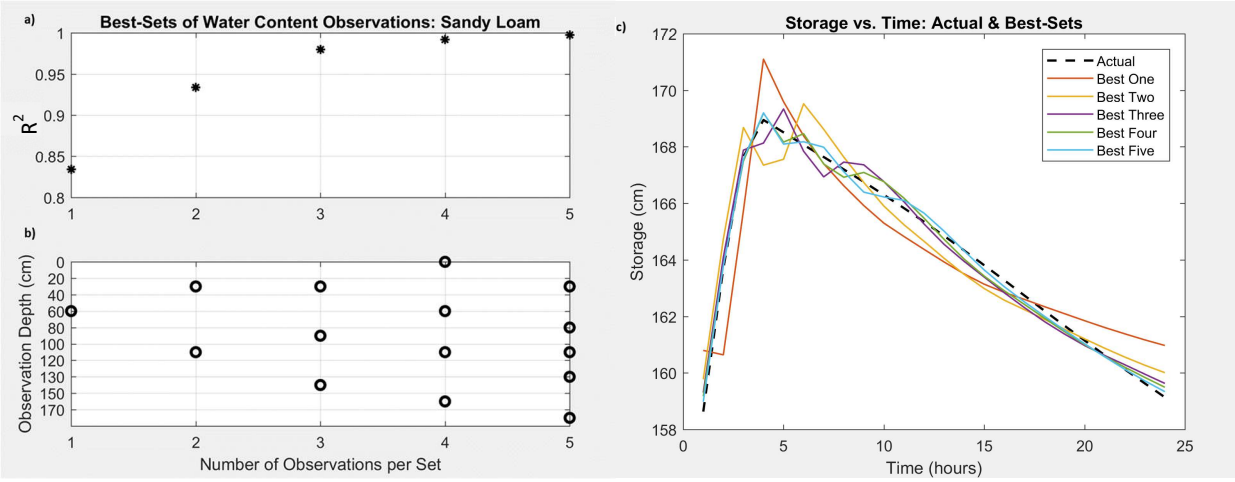

Figure 5: The observation sets with the highest R^2 for quantifying total storage for a sandy loam. The a) R^2 of b) sets of 1-5 were analyzed and compared. c) The model storage values (actual) and total storage for each best set versus time.

One advantage of uMLR is that it provides a comparison among observation sets with the same number of observations and among sets with different numbers of sensors (*Figure 4*). As expected, there is a wide range in the quality of sets of the same size: the spread in R^2 values for any observation set size shows the value of selecting optimal sensor depths. There is also both diminishing return in adding more sensors to the optimal set and reduced variation among sets of the same size as the subset size increases. Practically, in this case, uMLR shows that there is only a slight increase in R^2 from three to five observations, indicating that there is little value in collecting more than three observations. A researcher can also see that the benefit of selecting optimal sensors for this problem is more important for smaller observation sets, demonstrating the trade-off between the number and quality of observations.

For a sandy loam, water content varies throughout the entire depth interval (Figure 1e). Therefore, the best sets of water content observations are generally well-spaced measurements as deep as 190 cm (Figure 5b). While this general guidance can be inferred based on the simple rules for sensor placement discussed above, uMLR allows for more specific recommendations. If only a single observation is

collected, the water content sensor is placed at a moderately shallow depth of 60 cm (Figure 5b). This location is within the range of greatest temporal water content variation and coincides with the maximum depth of the wetting front associated with the periodic flooding (Figure 1e). The optimal set of two observations (R^2= 0.9338) (Figure 5a) places sensors at 30 cm and 110 cm (Figure 5b), within the region of largest water content change, but spaced around the 60 cm depth selected for a single observation. This reflects the balance of choosing observations within the most dynamic region, while spacing them to minimize redundancy. The best-three set (R^2 = 0.9799) (Figure 5a) places sensors at 30, 90, and 150 cm (Figure 5b), again spreading the sensors to cover the depth interval of greatest variability while maintaining separation among the probes. Note that these results also highlight another advantage of uMLR. Namely, because all observations sets are analyzed, the best set of more than one observation is not required to include the best single observation. In contrast, a sequential analysis would have fixed the selection of the 60 cm probe and then selected depths to add, thereby misidentifying the optimal sets for multiple observations.

In addition to selecting the optimal sensor depths, uMLR indicates how these observations should be weighted. For example, consider the best set of three sensors (30, 90, and 150 cm). The standard approach is to assume that each sensor represents the region in which it is placed. Then, the total length of water over the shallowest 200 cm would be calculated by applying the following weights on each measurement (90+30)/2-0; (150+90)/2 – (90+30)/2; and 200-(90-30)/2. That is, the total water stored is 60obs1 + 60obs2 + 80obs3. In contrast, the constants provided by the uMLR analysis define the total water storage as 117obs1 + 56obs2 + 46obs3 + 47 for sensors at these same depths. The deepest observation has the lowest weight, even though it 'covers' a larger region (it is farthest from the closest probe). This suggests that uMLR is accounting for both the spatial weighting and the information content of each observation, which would be impossible to intuit without using uMLR or some other, similar mapping approach. The total storage values calculated through time using the weighting provided by uMLR are shown in Figure 5c. The single best observation at 60 cm (Figure 5b) accurately estimates the timing of the maximum total storage but overestimates the magnitude of the maximum total storage (Figure 5c). As observations are added, the total changes in the inferred length of water stored through time are captured more accurately, with diminishing return after three observations are included.

Optimizing water content observations - sandy loam, measurement uncertainty

It is expected that measurement error decreases the correlation between observations and the outcome of interest. This should increase the number of observations needed to achieve a given predictive power and could alter the observations selected into the best set. To examine the impact of measurement error on the optimal observations sets, we repeated the previous analyses with added measurement error. Error was modeled as Gaussian with zero mean and a standard deviation of ±0.02 m^3m^{-3} based on the resolution of time-domain reflectometry measurements provided by Topp et al. (1980). Ten error realizations were created and uMLR was applied to each independently.

The uMLR results for sandy loam with measurement error are shown in Figure 4b. For this figure, the R^2 for each observation set was found using RDM across all error realizations. As expected, the maximum achievable R^2 for any set size was lower when measurement error was included (Figure 4b) than for noise-free data (*Figure 4*a). However, even with the modeled measurement error, we concluded that sets of three observations had sufficiently high R^2 values. It is likely that other such analyses would lead

an investigator to include more observations to account for the expected level of measurement error. A similar analysis could be conducted to examine sensors with different measurement uncertainties to compare the trade-off between using a few higher quality sensors versus more, less expensive sensors with higher measurement error.

The composition of the best sets is also affected by the addition of measurement noise. Considering measurement sets comprised of three observations (Figure 6), the recommended best sets are similar across error realizations, but there were small differences in the optimal depths. Like the error-free best set of three observations (dashed red lines; Figure 6), all of the best-three sets include shallow, middle, and deep measurements that are evenly distributed throughout the depth interval. The addition of measurement error consistently resulted in the placement of sensors at depths less than or equal to the depths with no measurement error. Specifically, 70% of the error realizations suggest 30 cm as the shallowest measurement, which is the same as for the noise-free data. In general, there is a smaller spacing between sensors for data with error, possibly reflecting the possible advantage of partially redundant data for noise cancellation.

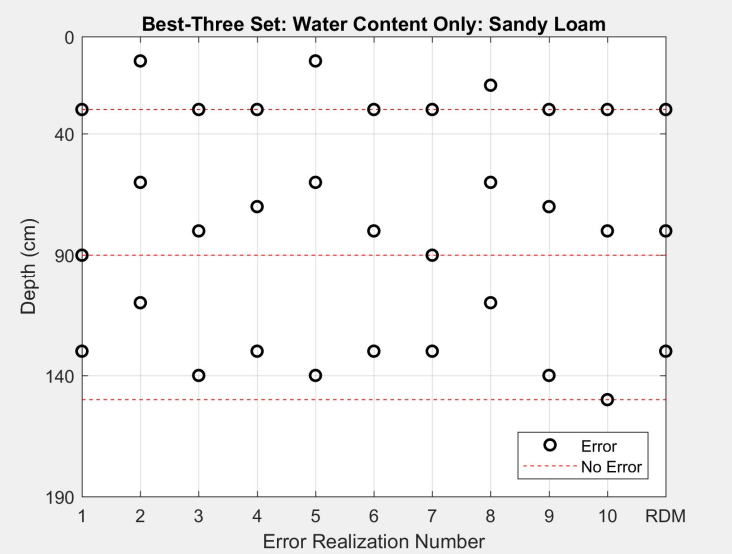

Figure 6: The depths of the best-three observations for error-free water content data (red dashed line) and for data with error added (black circles). The RDM best set for all 10 error realizations is included on the right side of the figure.

The RDM best set for a sandy loam has observations at 30, 80, and 130 cm (rightmost circles, Figure 6). This best set suggests relatively evenly distributed sensors that are slightly shallower than recommended for the error-free case. This may reflect reduced redundancy of closely-spaced observations when measurement error is considered. The R^2 for RDM set (0.9192) is only slightly lower than the maximum R^2 for each individual realization (0.9440-0.9629). But, it does indicate a further loss in performance for any single realization at the expense of risking a very poor performance for any anticipated conditions.

Optimizing water content and temperature observations - sandy loam, measurement uncertainty

The value of water content measurements is clear, but it is more difficult to predict whether temperature measurements would be worth including. To assess the value of adding temperature

measurements, we considered the water content measurements with error, as described above, and temperature measurements with error modeled as Gaussian with zero mean and a standard deviation of ±0.1 °C.

The best-three set with no error (dashed red lines; Figure 7) had two shallow temperature measurements (30 cm and 50 cm) and one shallow water content observation (30 cm). This was a surprising result, given the indirect relationship between temperature and water storage; it is highly unlikely that this result would have been proposed based on simple rules for sensor placement. One explanation may be that temperature fluctuations in the shallow subsurface are related to the flux through the root zone (Rousseau et al., 1999; Constantz et al., 2003; Shan and Bodvarsson, 2004), which relates directly to storage dynamics. Looking at the results more closely, uMLR indicates that two observations (one water content at 30 cm and one temperature measurement at 50 cm) result in an R^2 of 0.9959 which is not significantly different than the R^2 (0.9990) achieved with the best-three set. Demonstrating that the shallow temperature observation is largely redundant.

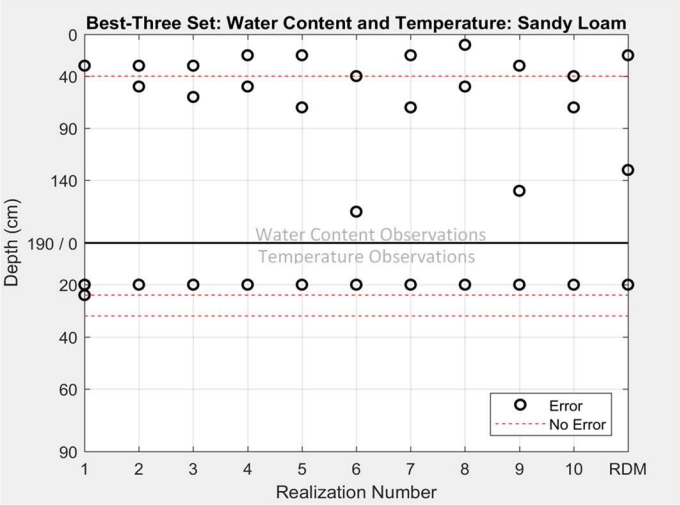

Figure 7: The depth of the best-three temperature and water content observations for observations with (black circles) with without (dashed red lines) error added. Ten error realizations are shown and the RDM best set is shown on the right.

Ten error realizations were considered, as previously for water content observations only (black circles, Figure 6). The best-three sets considering measurement error (black circles, Figure 7) show some variation in the specific sensor depths across the realizations. But, uMLR almost uniformly selects two (generally shallow) water content observations and one shallow temperature observation. Comparing the results with and without measurement error (black circles and dashed red lines, respectively; Figure 7), uMLR seems to favor a second water content observation to reduce the effects of water content measurement error at the expense of the largely redundant second temperature observation selected when observations had no error.

The RDM best set is similar to the individual error realizations with two water content observations (20 cm and 130cm) and one temperature observation (20 cm). However, the water content observations are spaced further apart for the RDM best set to collect information about the full range of water content variation that occurs throughout the soil column (Figure 1e). The R^2 for the RDM best set, using both

temperature and water content observations (0.9792), is higher than the RDM best set using only water content observations (R^2 = 0.9192) (Figure 6). The higher R^2 for the measurement design, using both temperature and water content sensors, demonstrates the usefulness of uMLR to design more effective measurement networks (with multiple sensor types) that might otherwise be impossible to intuit.

Optimizing water content and temperature observations – each soil, measurement uncertainty

We repeated the best set and RDM analysis presented above (both measurement types, with and without error) for each soil texture: clay; silt; silty loam; loam; and sandy loam. A best set of three observations was chosen for each soil for measurements with and without error (black circles and red points, respectively; Figure 8b) assuming in each case that the soil properties were known.

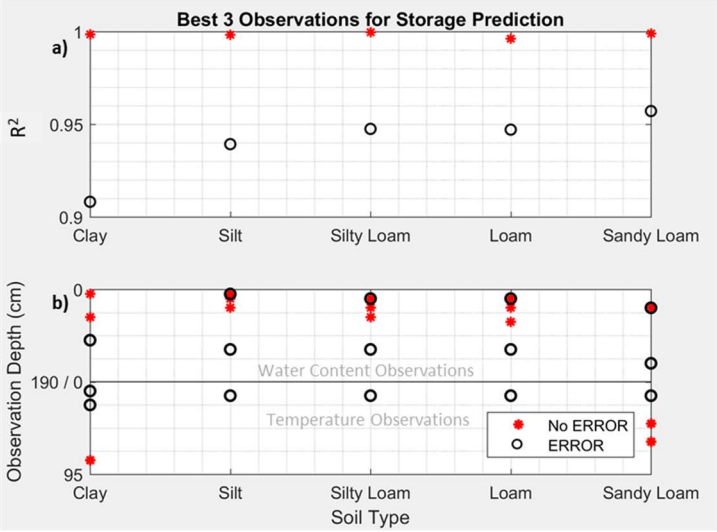

Figure 8: The best-three for the assessment of total hourly storage using uMLR for no error (red points) and RDM selected sets with error (black circles). Observation depths were suggested for individual soils (Clay-Sandy Loam). The top plot shows the R^2 and the bottom shows the observation depths for the best-three for each condition. The y-axis is divided into temperature and water content observation depth with an increase for each type indicating increasing depth (temperature 0 to 95cm and water content 0 to 190cm).

With no measurement error, for combined measurement types, uMLR primarily selected shallow water content observations (red points, Figure 8b). This is consistent with the intuitive choice of water content measurements placed at depths that show relatively large changes in time. The dynamic range varies among soil types (Figures 1); as a result, silty loam places the best-three observations at 10 cm, 30 cm, and 50 cm; silt observations were slightly shallower (0 cm, 10 cm, and 30 cm); and loam observations were slightly deeper (10 cm, 30 cm, and 60 cm). There were very few temperature observations included in the best sets (red points, Figure 8b). Interestingly, the exceptions are the coarsest and finest textures: sandy loam and clay. No single measurement design was selected consistently across the soil types.

The RDM analysis for each soil texture (black circles, Figure 8) primarily suggested two water content measurements and one shallow temperature measurement, similar to the sandy loam (black circles, Figure 7). The two water content measurements are spaced to gather information about both the

shallow dynamic zone and the deeper zone with less temporal variation. Coarser-grained soils have deeper observations with wider spacing because the dynamic region extends more deeply in coarser materials (Figure 1e). As discussed above, the shallow temperature measurement may provide additional information about the flux through the root zone. No single set was identified as optimal for all soil types, but the designs are more similar (black circles, Figure 8b) than those selected for the error-free conditions (red points, Figure 8b).

Optimizing water content and temperature observations – all soils, soil uncertainity and measurement uncertainty

Finally, we consider the selection of observations for assessing total water storage where the soil type/condition is unknown prior to installation and the observations have measurement error. Figure 9b compares the RDM best sets of three observations for each soil, considered independently (black circles) and for all soils, treating the soil texture as uncertain (blue lines).

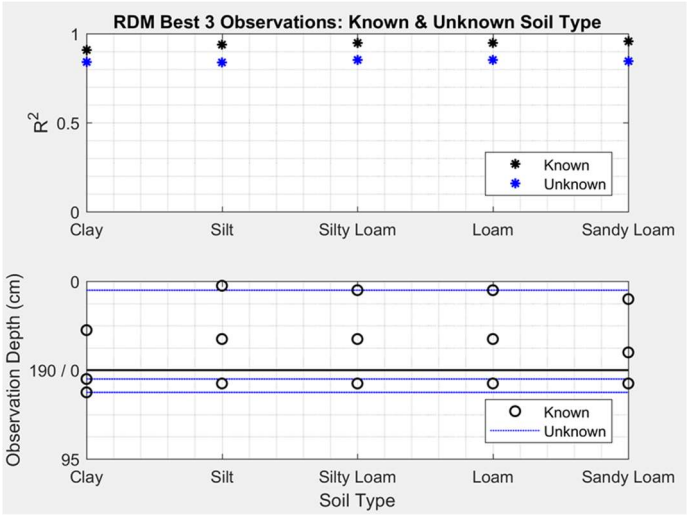

Figure 9: The RDM best-three observations for individual soils (black) and an unknown soil (blue). The top plot (a) shows R^2 values for both conditions and the bottom plot (b) shows the observation depths for the individual soils (black circles) and for the unknown soil (blue lines). The y-axis for the bottom plot is divided into temperature and water content observation depth (temperature 0 to 95 cm and water content 0 to 190 cm).

The best-three observations for individual soils (black circles, Figure 8; Figure 9) are similar across soil types which suggests that soil parametric uncertainty may not have a strong influence on the network design. If this is true, then we would expect that uMLR (with RDM) would select a single compromise design that has an R^2 close to those found for each soil treating its properties as known. However, when we applied the RDM approach, to consider both measurement error (10 error realizations) and uncertainty regarding the soil properties (5 soils), the RDM approach led to a best set of one shallow water content (10 cm) and two temperature observations (10 cm and 40 cm) (blue lines, Figure 9b) with an R^2 between 0.8530 and 0.8391 (blue points, Figure 9a). The minimum R^2 (0.8391) is considered to be the goodness of fit over all error realizations and soils (Figure 9a). The RDM design (blue lines, Figure 9d) is not optimal for any individual soil type (Figure 8; Figure 9b). Rather, it provides a unique solution that

represents a good compromise for all conditions considered: it seeks optimality while guarding against the expected sources of uncertainty.

The RDM approach is very sensitive to protecting against poor performance. For this application, the design most closely resembles the individual design for clay (Figure 8; Figure 9b). In fact, it appears that the RDM optimal design places considerable emphasis on discriminating between clay and other soil types. This conclusion is supported by repeating the analysis with clay removed from consideration: the RDM design for this condition more closely resemble the designs for the other individual soils with error (Figure 8); two water content observations at 10 cm and 160 cm and a temperature at 20 cm and having an $R^2 = 0.85$ (not shown).

A small amount of information about the uncertainty of the conditions (here, soil texture) can have important implications for the optimal, robust measurement design. In the example above, by simply restricting the soil type to exclude clay, the RDM best set design most closely resembles that of individual soils (silt-sandy loam) and the goodness of fit increases from 0.83 to 0.85. This type of analysis could provide useful information regarding ancillary information to collect that is likely to impact the measurement network design. In this case, it could lead a researcher to search for evidence that the finest soil types could be excluded from consideration. Most importantly, the RDM optimal design, with one very shallow water content measurement and two relatively shallow temperature measurements, would be difficult if not impossible to intuit without an objective optimization approach like uMLR with RDM. Furthermore, the flexibility of the uMLR with RDM method to assess tradeoffs in the measurement network design with little computational effort suggest that this approach could have wide applications in experimental design.

Conclusion

The uMLR with RDM approach is designed to quantify the benefits of different sets of candidate observations to infer specific experimental outcomes under user-specified sources of uncertainty. Using this tool, an investigator can choose the optimal number of observations needed to meet a measurement objective within a given budget. They can also examine the tradeoff between sensor quality and number of sensors. They can examine the value of ancillary information through consideration of the impacts of known uncertainties on measurement network design and performance. Finally, they can determine the optimal mix of sensors of different types and the correct placement of sensors for time-lapse monitoring. The key feature of uMLR with RDM is that it has low computational cost, allowing it to be applied to any study that will be evaluated in the context of a model. Despite its potential benefits, it should be noted that uMLR with RDM is not intended to replace more complete, formal data worth analyses. However, for more complicated applications, it may be worthwhile to initially apply uMLR and RDM to reduce measurement sets prior to implementing more computationally-costly data worth methods.

Chapter 2

Robust predictive design of field measurements for evapotranspiration barriers using universal multiple linear regression

Melissa Clutter, Ty P.A. Ferré, Zhuanfang Fred Zhang and Hoshin Gupta

In Submission: Water Resources Research

Abstract

Surface barriers are commonly installed to reduce downward water movement into contaminated zones. Specifically, evapotranspiration (ET) barriers are used to store water and release it, via ET, before it can percolate into an underlying waste zone. To assess the effectiveness of a surface barrier, neutron probe or other types of sensors may be used at different spatial locations to monitor this store-and-release mechanism. We used an existing data set, model-simulated data, and a dimensionality reduction approach called universal multiple linear regression (uMLR), to optimize the required number of sensors in a 2-m thick surface barrier. To understand the usefulness of implementing predictive uMLR prior to sensor installation, we compare several network designs, selected based on down-sampling of existing data, with a recommended sensor design based on model-simulations performed without consideration of existing data. We found that uMLR, combined with robust decision making, provides a simple and high-quality network design for robust monitoring of the total water stored in a surface barrier.

Introduction

Evapotranspiration Barriers

For over 30 years, the US Department of Energy has considered surface barriers to be an integral part of the remediation process for contaminated zones (USDOE, 1987). Evapotranspiration (ET) barriers are a type of surface barrier used to reduce (or eliminate) the movement of water into an underlying waste zone. ET barriers rely on two mechanisms to reduce downward water movement to depths where it could mobilize contaminants: 1) storage of water in the soil; and 2) release of water via ET. The integration of these two processes is referred to as store-and-release.

Store-and-release mechanisms are particularly important in arid and semi-arid systems, where the potential evapotranspiration (PET) exceeds precipitation (Wilcox et al., 2003; Zhang, 2016). For example, the average recharge beneath undisturbed natural semi-arid vegetation may be less than 5 mm/year due to store-and-release processes (Fayer and Keller, 2007; Zhang 2016). However, without vegetation, drainage can be as high as 50-100 mm/year (Scanlon et al., 2005; Zhang 2016). Consequently, an effective ET barrier can nearly eliminate drainage in arid or semi-arid regions.

To ensure the success of ET barriers in preventing contaminant mobilization, researchers monitor the store-and-release mechanisms by using sensors such as neutron probes (e.g. Zhang, 2015, 2016). One successful monitoring approach is to install several access tubes and then conduct measurement campaigns where the probe is lowered to a series of depths, neutrons are counted over an appropriate averaging time, and the procedure is repeated at the next access tube (Gardner and Kirkham, 1952; Zhang, 2016). Each measurement, calibrated to water content, is assumed to represent a depth interval, allowing for a conceptually simple calculation of the total water stored in the barrier. Neutron probes

offer high depth resolution, but because measurements are manual and time consuming, observations are often collected with low temporal resolution.

An alternative is to install sensors, such as time-domain reflectometry (TDR) probes at multiple depths (Topp et al., 1980). These probes can be interrogated automatically, allowing researchers to capture more dynamic conditions. But cost, and concerns about disturbance of the subsurface, generally limit the number of probes that can be buried at different depths at each location, and point measurements commonly offer high temporal resolution with reduced depth resolution.

Finally, a hybrid strategy is considered here: infrequent high depth resolution and high temporal monitoring with neutron probes to understand the details of water movement, combined with a small number of buried TDR probes to monitor total water storage. Our goal is to develop a generalizable method that can be used to determine the necessary number of point measurements, and the optimal depths for those sensors, to provide an acceptably good measurement of total water storage through time.

Focus and Scope

Budget limitations often restrict monitoring network design and are a common challenge for data collection (Hubbart et al., 2005; James and Gorelick, 1994; Liu et al., 2012; Mogheir et al., 2003). Therefore, measurement network designs often rely on some form of data worth analysis prior to sensor installation. Data worth analyses effectively allocate a sampling budget by identifying the most valuable data for measuring the outcome of interest, given the current level of understanding. Data worth analyses include formal techniques that use quantitative metrics to identify the most valuable data, and informal methods that rely on intuition or trial-and-error (James and Gorelick, 1994; Strebelle, 2002; Mogheir et al., 2003; Tiedeman et al., 2003; Çamdevýren et al., 2005; Fernandez-Galvez et al., 2006; Vereecken et al, 2008; Jolliffe, 2011; Sun et al., 2013; Kikuchi, 2017; Lin et al., 2017). Informal and formal methods are complementary with a joint goal of identifying field measurements to be both informative and efficient. This goal must be balanced with the effort needed to complete the monitoring network design optimization. Some experimental efforts justify the use of more complex and computationally expensive data worth analyses (Strebelle, 2002; Hermans et al., 2015; Kikuchi, 2017, Ju et al., 2018, Vilhelmsen and Ferre, 2018). But, there are also relatively simple problems that can benefit from less computationally complex optimization approaches (Ju et al., 2018). Furthermore, there may be a role for simpler analyses to help guide and justify more complex data worth calculations.

For this study, we use universal multiple linear regression (uMLR) (Norgaard et al., 2014; Clutter and Ferre, in press). Previously, Norgaard et al., (2014) used this approach with pre-existing data to identify a reduced set of parameters that could be measured as a surrogate for performing more costly colloid mobilization and transport analyses. Clutter and Ferre (in press) expanded the method to better address uncertainty in user-defined predictions of interest, by introducing a predictive approach that combines physically-based model results, uMLR, and Robust Decision Making (RDM) (Lempert and Groves, 2010; Lempert et al., 2013) to optimize network design prior to data collection; they applied the approach to a relatively simple hypothetical system including a stream with cyclic boundary conditions.

Here, we apply uMLR to optimize sensor installation based on an existing field data set. This study is meant to be a proof of concept using existing field data. However, it is not intended to supply a specific network design for the site considered. We first show that uMLR with RDM (hereafter uMLR-RDM) can

be used to hindcast what would have been optimal sensor placement, by down-sampling previously collected field data from the Prototype Hanford Barrier (PHB). We then show that the predictive uMLR-RDM approach could have been used to design a simple monitoring network prior to data collection. Finally, we show how uMLR-RDM results can be used to define general, practical rules for monitoring network design; this would not be possible with variable reduction methods designed to define a single best network design.

Site Description

The PHB, an ET cover system, was constructed in 1994 over an existing contamination area (DOE-RL, 2016; Zhang, 2016). The Hanford Site is a semi-arid zone with hot, dry summer and cool, wet winters (Hoitink et al., 2005), and the PHB design is aimed at eliminating drainage during all potential seasonal variations in water content over a 1000-year period (Zhang, 2016). The primary layer that processes store-and-release of the PHB system is a 2-m vegetated silt loam storage layer. The vegetation is composed of semi-arid grasses and shrubs (Rickard and Vaughan, 1988) with a uniform plant root density to 1.1-1.5 m depth (Gee et al. 1995, 1996).

From 1994-2013, the water content was monitored at 12 stations; six stations in the northern section and six in the southern section, spaced 5-m apart (Figure 1). A comprehensive summary of the performance of the PHB from 1994 to 2013 is given in DOE-RL (2016). The monitoring data are described in the Appendix R of DOE-RL (2016). Zhang (2015) analyzed the field water retention of the silt loam layer at 4 depths at each of the 12 access tube stations. The soil is considered nearly uniform across all 12 stations and the non-hysteretic average van Genuchten parameters are θ_s = 0.344 ± 0.056 (mm^3mm^{-3}), θ_r = 0.068 ± 0.012 (mm^3mm^{-3}), α = 5.46 x 10^{-3} (mm^{-1}), n = 1.51 ± 0.13 (-), and K_s = 4.49 x 10^4 (mm/year) (Zhang 2015; 2016). At each station, counts were measured at 0.15 m intervals using neutron probes deployed in vertical aluminum access tubes extending to 1.9 m depths (13 total depths). Counts were collected manually once per month during a typical water year and 2-5 times per month during a wet year. The water content was then calculated using (Zhang, 2015)*:*

$$\theta = a_0 + a_1 N + a_2 N^2 \quad (1)$$

where θ is the water content (m^3m^{-3}), a_0 = -0.01649, a_1 = 1.449 x 10^{-5}, a_2 = 3.234 x 10^{-10}, and N is the 16 second neutron count.

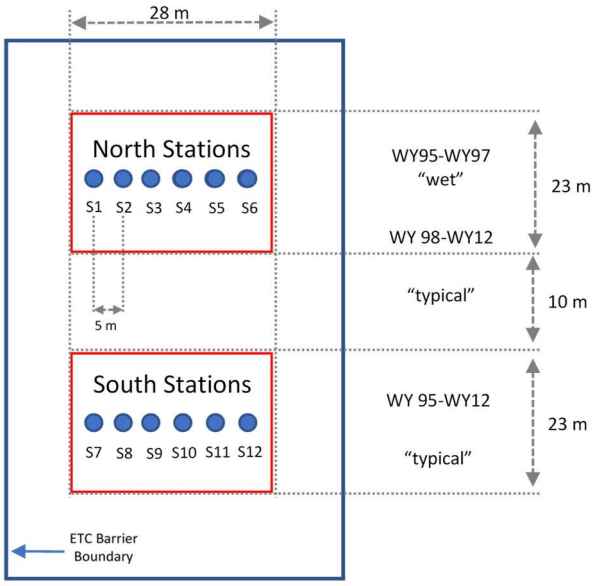

Figure 1: PHB study site. 12 stations where 1.9 vertical access tubes were installed to 1.9 m depths and spaced 5 m apart. The six northern stations experienced wet conditions during WY95-WY97 and typical conditions WY98-WY12. The six southern stations experienced typical conditions during the entire data collection period.

A water year (WY) is a 12-month period from October 1st of the previous year to September 30th of any given year (e.g. the water year ending September 30th, 2013 is WY13). From WY95 to WY97, an enhanced precipitation test was carried out on the northern section (Figure 2: blue). Irrigation was added to natural precipitation events to increase the maximum precipitation stress on the barrier to represent a 1000-year precipitation maximum (annual precipitation at the 99.9% percentile). Note, we refer to this period as "wet" (WY95-WY97) and the remaining years are referred to as "typical" (WY98-WY12) (Figure 2). Only the northern stations experienced wet conditions and all stations experienced typical precipitation events during the water years specified. During wet years, an average runoff of only 2 mm/year was observed (Zhang, 2016); this low runoff can be attributed to the high saturated hydraulic conductivity of the silt loam (Gee et al., 1995). The ET processes varied throughout the year but were strong enough to release almost of all the stored water in the barrier (Zhang, 2016). The average drainage rate through the barrier was ~0.005 mm/year; two orders of magnitude below the maximum design drainage rate of 0.5 mm/year (Zhang, 2016).

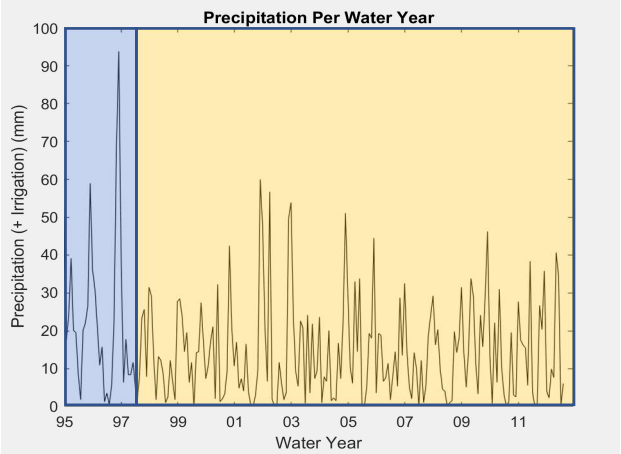

Figure 2: Monthly precipitation for the 18-year study period at the northern stations. Wet years (WY95-WY97) have additional irrigation water added to simulate a 1000-year precipitation event (blue). Typical precipitation years are shown in yellow. The southern stations experienced typical precipitation during WY95-WY97 (not shown).

Water content profiles at the PHB were similar across all 12 stations and differed between wet and typical years. Representative profiles (here shown for station 1) are presented in Figure 3 for both conditions. Under wet conditions, the wetted zone (>0.2 m^3m^{-3}) in all six northern stations extended to depths of at least 180 cm (Figure 3a). During typical years, the wetted zone (>0.15 m^3m^{-3}) only extended to approximately 75 cm (Figure 3b). Total water stored is calculated by summing the products of the water content at each depth and the measurement interval. The total water stored varied from 400-700 mm during wet years and 50-100 mm during typical years (Figure 3c) and was similar across all stations receiving the same treatment.

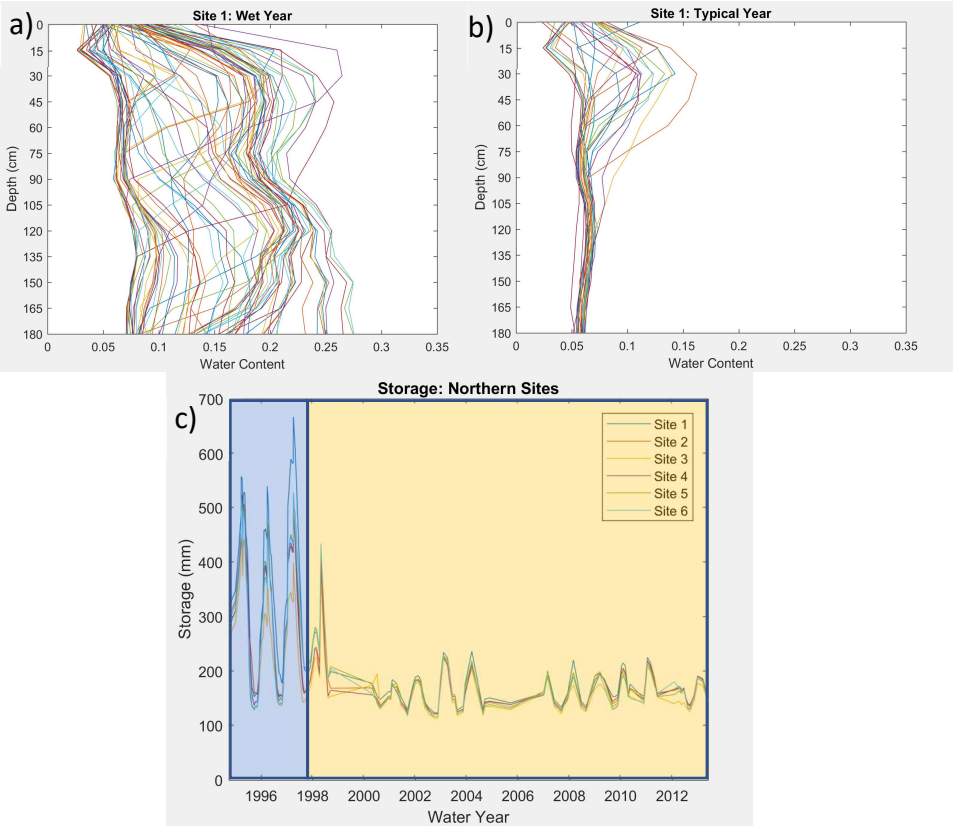

Figure 3: Northern stations (Figure 1, stations 1-6) experience both wet and typical precipitation conditions (see Figure 2). The measured water content versus depth for station 1 for wet years (WY95-WY97) and typical years (WY01-WY03) are shown in a and b, respectively; we only show station 1 because the profiles vary only slightly across stations. Each line represents a different time at which the neutron profile was logged; data was collected more frequently during wet years. c) shows the total water stored for northern stations 1-6 during wet years (blue) and typical years (yellow).

Methodology

Universal multiple linear regression is a computationally inexpensive approach that assumes that combinations of observations (water content at different depths, θ_i) can be linearly related to a prediction of interest (total water stored, W).

$$W = a_0 + \sum_{i=1}^{n} a_i \theta_i \qquad (2)$$

Where a_0 is the y-intercept, a_i is the regression coefficients, and n is the number of observation depths.

The uMLR method does not assume that the original processes were linear. Rather, it simply assumes that the goodness of fit of a linear regression (R^2) is a measure of the information content of the measurement set that is relevant to the prediction of interest.

The computation efficiency afforded by the linear assumption underlying uMLR allows us to consider all possible combinations of observation depths, up to any given subset size, n (Equation 3). The goal of uMLR is not solely to identify a single optimal measurement set. Rather, it quantifies the likely value of *all* unique combinations of different observation depths. By analyzing these results, we can derive

general rules for designing optimal observations sets and gain insight into the value of proposed observations. These analyses, performed over the ensemble of measurement sets, is the central strength of uMLR.

Because every possible observation set can be analyzed independently, uMLR is suitable for parallelization when many sets must be analyzed. However, the large number of possible combinations that can be formed can limit the number of sensors to be considered. Specifically, the potential number of observation sets (p_{set}) is given by:

$$p_{set} = \sum_{n=1}^{s_{max}} \frac{c_o!}{n! * (c_o - n)!} \quad (3)$$

where c_o is the number of candidate observations, and s_{max} is the maximum observation set size (Clutter and Ferre, in press). Therefore, some judgement is required to define the candidate observations and the maximum set size before applying uMLR to confirm an appropriate run time.

Down-sampling of existing PHB data

There are many factors that go into network design, and for research efforts there are often reasons to collect more data than that which is strictly necessary. However, for more practical applications there might be reasons to reduce the number of sensors. For down-sampling, the uMLR approach aims to answer the question: *with the benefit of hindsight, could the total storage have been quantified at each measurement time using fewer sensors, appropriately placed and weighted?* We use the PHB data set to demonstrate the down-sampling approach. We do not intend to make recommendations for a simpler network design for the PHB site, specifically. Rather the data are used as a proof of concept for other applications that may be subject to stricter budgetary limits.

The PHB was designed to be nearly uniform. But, to demonstrate how the method can be used in natural conditions, we treat the results as if the optimal set is expected to vary among the access tube locations. Therefore, we investigate each individual access tube location (Figure 1; stations 1-12). According to Equation 3, if we use 13 candidate observations (c_o) (at each station) and limit the down-sampling to 5 or fewer sensors (s_{max}), there are 2,379 potential observation sets (p_{set}): 13 sets consisting of 1 observation, 78 sets of 2 observations, 286 sets of 3 observations, 715 sets of 4 observations, and 1,287 sets of 5 observations.

We first used uMLR to determine the number of sensors needed. We next evaluated all possible measurement sets of a certain set size (e.g. sets of 3) for each location (12 stations). Finally, we found the minimum R^2 for each observation set across all stations and assigned this R^2 value to the observation set. From these results, we identified the '*RDM Best Set*' consisting of all of the observation sets (e.g. sets of 3) that had an acceptably high R^2 *at all stations*. This minimax strategy is referred to as '*Robust Decision Making'*; for more details, see Clutter and Ferré (2019).

Predictive approach using simulated data

Using simulated data, we can apply uMLR-RDM to answer the question: *could an optimal sensor network have been designed for a surface barrier prior to data collection?* This requires four steps: 1) develop a forward flow model; 2) identify sources of uncertainty; 3) create an ensemble of models to simulate candidate observations and total water stored considering all sources of uncertainty; and 4)

apply uMLR-RDM to select a network that would be expected to perform satisfactorily over the ensemble (Clutter and Ferré, 2019).

In this study, we used HYDRUS 1D (Šimůnek et al., 2013) to generate both the outcomes of interest and the candidate observation sets in response to site-specific boundary conditions (Zhang, 2015, 2016; Hoitink et al., 2005; Gee et al. 1995, 1996; Rickard and Vaughan, 1988). We simulated responses to the measured initial and boundary conditions reported in Zhang (2015; 2016) for wet and typical conditions; however, additional simulations could have considered precipitation uncertainty as well. For both wet and typical conditions we considered a range of five soil textures (including silt, loam, sand and two others; see Table 1) to form an ensemble of 10 models. We primarily consider uncertainty in boundary conditions and soil hydraulic parameters. However, if we had considered more sources of uncertainty (e.g. climate, barrier thickness, vegetation parameters, etc), the model ensemble would have been larger, but the general procedure would not have differed.

For down-sampling, we found an observation set that performed acceptably well ($R^2 > 0.90$) for 12 different stations using uMLR-RDM. In a post-audit sense, this identified a network that can be expected to perform acceptably well over the range of hydrogeologic conditions at the 12 stations. For the predictive design case, we limited HYDRUS to a design for one station, but subjected the model to a wider range of soil hydraulic parameters. In the RDM context, this represents consideration of the effects of differences in hydraulic properties among well locations, while still assuming that the medium is locally homogeneous. Candidate observations were located every 0.15 m throughout the 2 m column depth. Monthly water content values were simulated at the candidate depths (Figure 4; Figure 5) and the total water storage value was calculated every month (Figure 6).

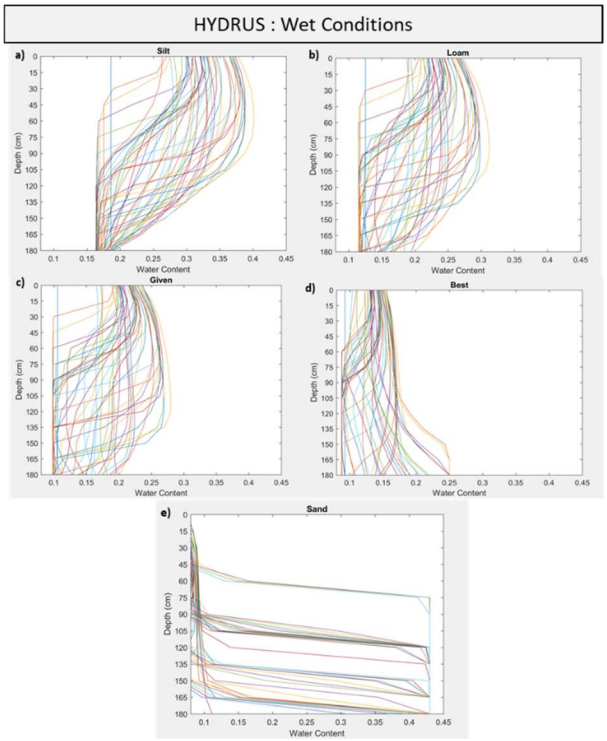

Figure 4: Simulated water content vs. depth for all 5 soil types (Table 1) in the HYDRUS model ensemble during wet conditions (WY95-WY97). Each line represents monthly data.

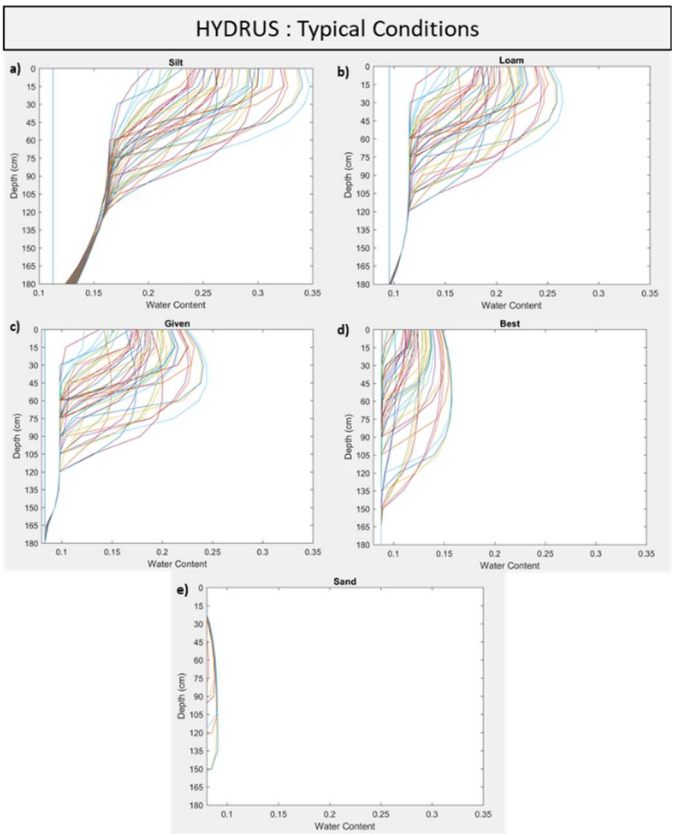

Figure 5: Simulated water content vs. depth for all five soil types (Table 1) in the HYDRUS model ensemble during a typical precipitation year (WY01-WY03). Each line represents monthly data.

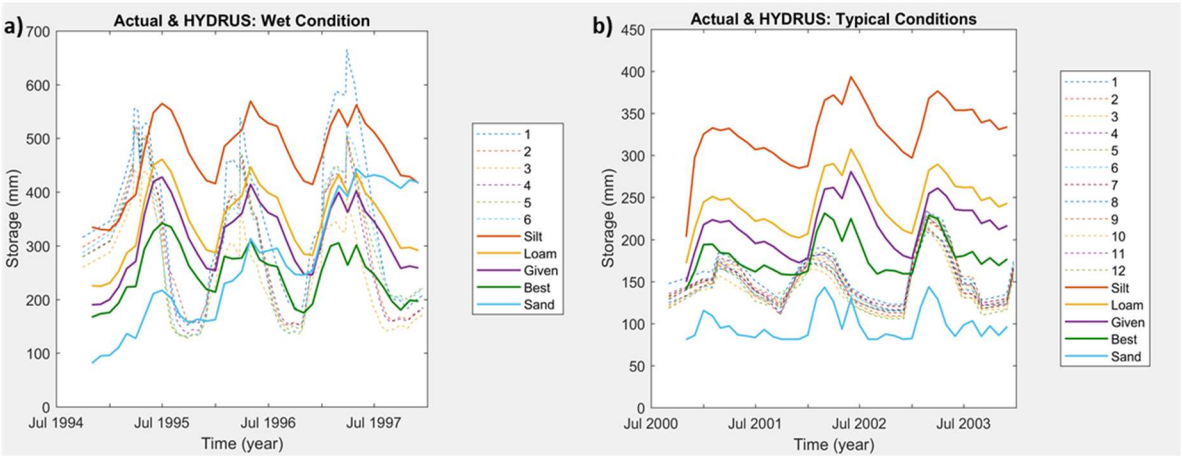

Figure 6: Simulated water content vs. depth for five simulated soils (see Table 1) in the HYDRUS model ensemble b) during a) wet precipitation conditions (WY95-WY97) and typical precipitation conditions (WY01-WY03) (bold lines). For comparison, the actual storage values (dotted lines) are also included for the monitoring stations at the PHB (see Figure 1).

We conducted this analysis for both wet and typical conditions, and then selected a design that would be acceptable for each condition. Note that this approach can easily be extended to consider other sources of uncertainty including model structure, soil layering, or plant root water uptake. We limited

our analyses so that we could assess how well the pre-screened network would have performed based on the existing (real) data set.

The HYDRUS model (with dynamic time stepping and an initial and minimum time step of 0.0001 years) was used to simulate water flow for 4 wet years and 15 typical years in a 2 m soil column. The top boundary condition was set to be constant flux, equal to the precipitation (mm/month) values reported by Zhang (2016) (Figure 2). The bottom boundary condition was set to zero flux (Zhang, 2016); however, uncertainty regarding this boundary condition also could have been examined as part of the model ensemble. The initial conditions were designed to minimize the spin-up time of the model by setting a pressure head that differed for the environmental conditions considered (-10 m for wet and -60 m for typical). For typical conditions, six years of average annual precipitation was used as spin up to minimize impacts of the initial conditions; for wet conditions, only one year of spin up was simulated and appeared to be sufficient.

The potential evapotranspiration rate was set to 500 mm/year during an average precipitation year and 600 mm/year during the 1000-year precipitation events. Using a built-in HYDRUS capability, the ET was set to be a constant value equal to 1% of the total daily value between 0-6 a.m. and 6 p.m.-midnight, and followed a sinusoidal pattern (with a maximum at 12 noon) during the rest of the day. Root water uptake was computed using the Feddes et al. (1976) model and the following parameters/conditions that are selected to be representative of typical desert shrubs (Kurc and Small, 2004; Sela and Assouline, 2015): anaeobiosis pressure head (ψ_m) of -54 mm and optimal uptake ψ_m between -224 mm and -326 mm; the water uptake decreases linearly between ψ_m of -326 mm and -8000 mm and uptake is equal to zero for ψ_m less than the wilting point of -15574 mm.

Five homogeneous soils were analyzed. The range of soils included were coarser and finer than the values reported in Zhang (2015) providing a range of plausible conditions; the soils that were too coarse or too fine to effectively store-and-release (as designed) were not considered. The hydraulic properties are described using the van Genuchten (1980) functions (Table 1). Default HYDRUS 1D soil hydraulic parameter values were used for three soils silt, loam, and sand. The other two soils types considered were 1) the local soil type at PHB, represented using manually calibrated parameters based on PHB storage values (labelled “Best”) and 2) the soil type represented using parameters reported by *Zhang* (2015) (labeled “Given”) (Figures 4-6).

Soil Parameters					
Soil Type	θ_s (-)	θ_r (-)	α (1/mm)	n (-)	K_s (mm/year)
Silt	0.034	0.46	1.60E-03	1.37	2.19E+04
Loam	0.078	0.43	3.50E-03	1.56	9.11E+04
Given	0.068	0.344	5.00E-03	1.51	4.49E+04
Best	0.068	0.25	5.00E-03	1.51	2.00E+05
Sand	0.045	0.43	1.45E-02	2.68	2.60E+06

Table 1: The van Genuchten soil parameters for the 5 soil types used in the study: residual water content (ϑ_r), saturated water content (ϑ_s), alpha (α), n, and saturated hydraulic conductivity (K_s).

From Figures 4-6 we can see that the water content profiles and total water stored are quite different among the soil types considered in the ensemble. The water content profiles for wet conditions have variations throughout the profile (Figure 4), while typical conditions have shallower variations (Figure 5). For both conditions, the ranges of water contents in the profile depend on the soil hydraulic properties of the soil (Table 1), and soils with similar soil hydraulic properties have comparable profiles. Generally, the volume of water stored-and-released is similar across soils, but the maximum/minimum volumes of water stored differs with soil type (Figure 6). For sand under wet conditions (Figure 4e; Figure 6a, blue), water builds up at 2 m and is not fully released via ET; we consider this a non-behavioral case according to the PHB engineering design and therefore did not use the sand results (under wet conditions) in our uMLR study.

Results and Discussion

uMLR and RDM: Down-sampling of existing PHB data

The water content shows measurable changes with depth and time (Figure 3a,b). However, water contents are also linked across space and time through infiltration and redistribution processes. Therefore, we can expect there to be considerable redundancy in the water content observations. Based on this, it should be possible to reduce the number of observation depths, with minimal loss of information, if the sensor depths are optimized. In general, this will require that sensors be placed at depths that show the highest temporal variability, and that sensors be located in such a way as to minimize redundant information.

Following these simple guidelines, based on the HYDRUS simulation results, we can expect that probes would need to be spaced throughout the 180 cm profile during wet years to capture all areas experiencing water content variation (Figure 3a). Meanwhile, during typical precipitation years, probes could be restricted to the shallowest 100 cm (Figure 3b). To maximize sensor separation, observations could be focused in a) shallower regions that are dominated by dynamic infiltration, drainage, and ET, and b) deeper regions that show slower, less extreme water content variations with time. Using this informal data worth analysis, we can develop a generic measurement design based on capturing the greatest variability in water content. However, while based in logic and understanding, this approach requires a considerable degree of subjective judgement. We instead applied uMLR to develop an objective set of recommendations for both the number and depths of sensors.

The results of an uMLR contain information about many observations sets. Figure 7 (left panel) shows the R^2 achieved for every possible set comprised of a given number of sensors. The best performing sets (highest R^2) are shown as black points, and the very high R^2 values indicate that the linear assumption underlying uMLR seems justified for this application. The large spread of R^2 values among sets for any given set size indicates that there is considerable value in carefully selecting the probe placement – in all cases shown, the choice of which depths to observe has a strong influence on the information content of the observations. Finally, the very minimal improvement (<0.01) achieved in the highest R^2 by considering more than three observations indicates that an acceptable network can be constructed with only three measurement depths. There is some level of user-preference in the decision of acceptable improvement with added observations. Note, a more formal, cost-based decision could be applied to determine the maximum number of sensors that could be justified to meet budgetary restrictions. The number of sensors will always represent a cost-benefit trade off and will differ among users.

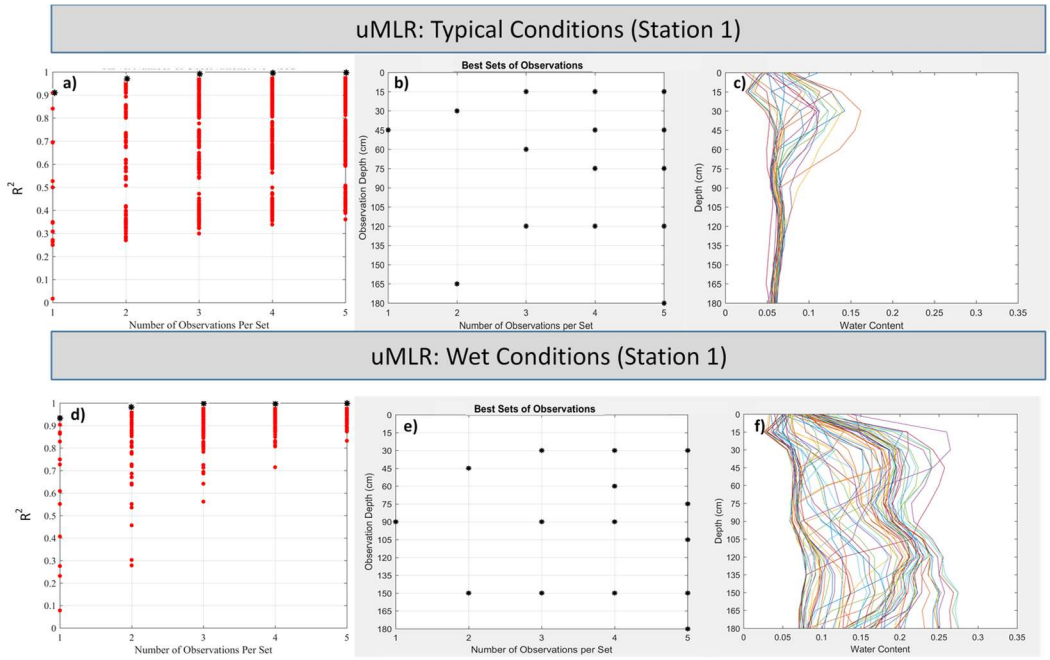

Figure 7: The uMLR approach for station 1 during typical (a,b,c) and wet conditions (d,e,f). The left panel (a,d) shows the R^2 for every combination of observation for sets of up to 5 sensors. The middle panel (b,e) shows the depths for the best sets of observations for set sizes 1-5. The right panel (c, f) shows the water content profile.

Total water stored can be calculated with acceptable accuracy using the best set of 3 observations (Figure 7). For example, for station 1 the regression coefficients a_0 through a_3 are 5.68, 523.22, 589.26, and 661.86, respectively, for the best three at depths 30, 90, and 150 cm (wet) and are 7.87, 273.72, 504.01, and 945.75 for the best three depths at 15, 60, 120 cm (typical). When uMLR is applied to an individual site, it returns the optimal sensor depths and the weight that should be applied to each to infer the length of water stored through time using Equation 2. The weights for all stations are included in Table 1 of the Supporting Information. Figure 8 shows the actual storage calculations using all 13 measurements (measured) and the storage calculated using only 3 observations (uMLR). The average water stored during wet years is 234 mm and 131 mm during typical years. The difference between the two storage values using the best set for station 1 is shown in black (measured – uMLR). The average absolute error is 2.85 mm, maximum absolute error 13.97 mm, and the RMSE is 3.86 mm during wet years; typical years are 6.86, 1.67, and 2.38 mm, respectively. The error values for all stations and conditions are given in Table S1 of the Supporting Information.

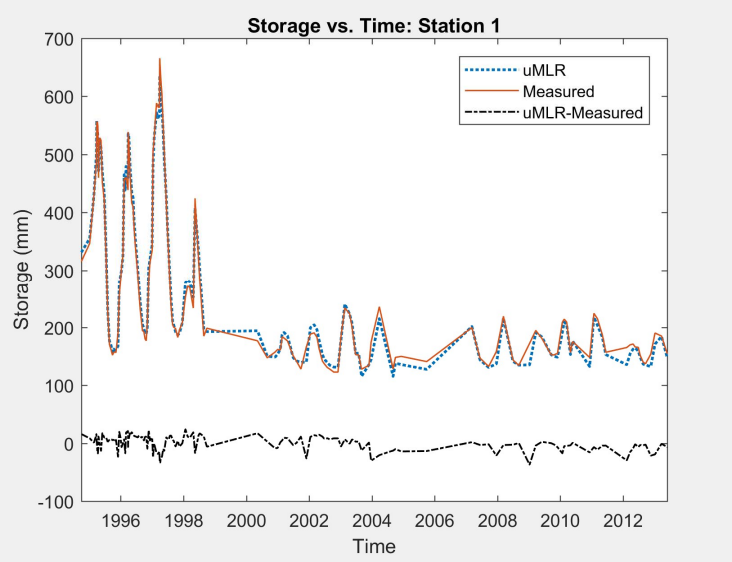

Figure 8: The storage through time using all measurements (red) and the best 3 observations appropriately weighted using Equation 2 (blue) for station 1. The difference between the two is shown at the bottom in black.

In addition to revealing the performance achievable using different observations sets, uMLR also retains the composition of each set. From this, a researcher can identify which sensors to include in a network as a function of the number of sensors included. As shown in Figure 7b, if only a single observation is to be selected, it will be at a depth where there is very high temporal variation of water content (as shown by Figure 7c). If two sensors are chosen, one should be near the bottom of the domain, to quantify the constant, deep water content, and the other at a shallow depth that shows high temporal variation. When a third sensor is added, a shallow sensor should be included to capture near surface water content variation. Interestingly, no single sensor depth is included in all three of these optimal measurement sets. In other words, uMLR identifies the set of observations that maximizes the total information that is relevant to the objective. Similar results are found for the wet conditions (Figure 7 d,e,f), but now the sensor placement is deeper because the water content varies to greater depth under wetter conditions. (Note: for wet conditions, only the northern stations are analyzed (stations 1-6); during typical conditions all stations are analyzed.)

The analysis of uMLR results (Figure 7) indicates minimal improvement in optimal set performance for more than three sensors. For clarity of presentation, the remainder of the study focuses on “best-three” sets. The compositions of the best set for each location, and for the RDM set across all locations, are shown in Figure 9. Although a common set of observations is identified, the weights that should be applied for these depths are still site specific (Table S2, Supporting Information). The RDM best sets (Figure 9a, b; RDM) for the different precipitation regimes reflect the differences in water content variability for each condition (Figure 4; Figure 5). For most access tube locations, during typical years, there are two observations located at depths <75 cm plus one deeper observation (Figure 9a; 1-12). The middle observations are closer to the surface and capture shallow changes in storage (Figure 5). During wet conditions (Figure 9b; 1-6), sets with the highest R^2 are evenly spaced because water content varies throughout the entire depth interval (Figure 4). The RDM best set for typical conditions is at depths 15,

60, and 165 cm (Figure 9a; RDM) ($R^2 \geq 0.97$) and for wet conditions is at depths 30, 75, and 135 cm (Figure 9b; RDM) ($R^2 \geq 0.98$).

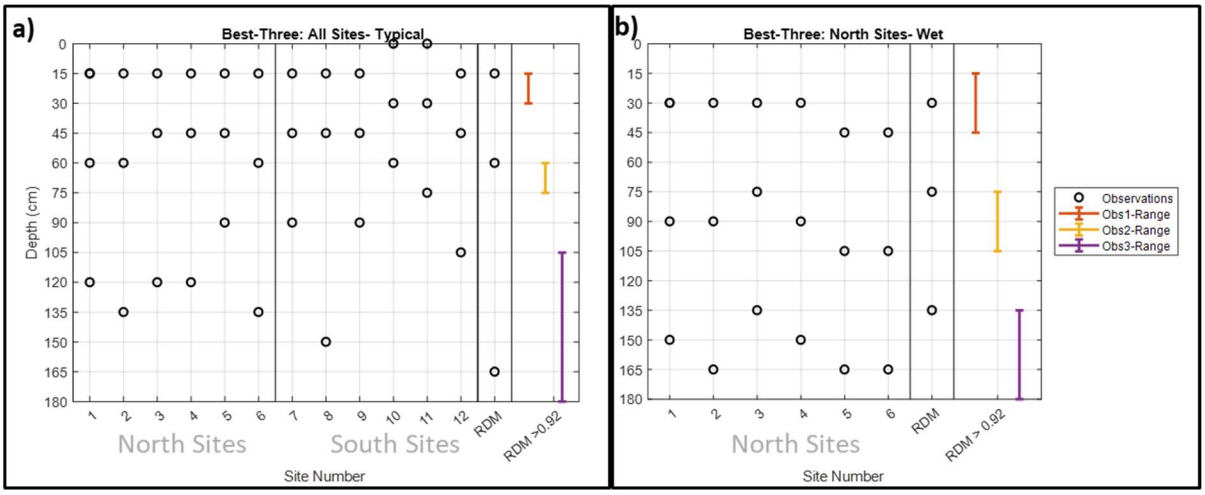

Figure 9: Each subplot shows the individual observation sets with the highest R^2 for each North and South station, and for the best RDM set across all stations for the down-sampling of existing data. Subplot a) is for typical conditions and b) for wet conditions. On the far-right side of each subplot the ranges of observation depths for which RDM $R^2 > 0.92$ are shown.

Importantly, robust measurement designs are not necessarily limited to the *single* best set from RDM (Figure 9a,b; RDM). Using uMLR, for both wet and typical conditions, we find ~100 combinations of observations that give an $R^2 > 0.90$ across all stations. Using these results, we can 1) implement RDM to find observations that perform well across a range of precipitation conditions and 2) identify a range of observation depths that give an acceptably high R^2 for each individual precipitation regime.

For the best set of observations that would work well for the entire range of precipitation conditions considered, we use the uMLR results from each individual condition (typical and wet) and then use RDM to find that observations that give an acceptably high R^2. The RDM best set (at depths 30, 90, and 150 cm) gives an R^2 of at least 0.90 for all stations and precipitation years. This measurement design is similar to the RDM design for wet years and could be used for uncertain precipitation conditions.

For a depth interval analysis, we can choose the minimum interval that we would want to consider for practical guidance on sensor placement, the minimum separation between the bottom of one measurement interval and the top of the next interval, and a threshold of acceptable R^2. Setting these to 15 cm, 30 cm, and 0.92, respectively, we obtain the results shown in Figure 9. Therefore, using the RDM approach, we can identify depth intervals in which any combination of three sensors that are installed within the specified depth intervals will have an acceptably high R^2. Our results indicated seven combinations of three depth intervals that satisfy these conditions. From these, we selected the set covering the largest range of depths, but other researchers may choose their preferred intervals for other practical reasons. The depth intervals for the wet and typical conditions are shown (Figure 9) as vertical colored bars spanning the three observation depth ranges. The two shallowest recommended observation ranges are narrower and shallower for the typical conditions, reflecting the restriction of water content changes to the near surface for typical conditions. Simultaneously, the deep observation is broader reflecting the near-constant water content with depth.

uMLR and RDM: Predictive approach using simulated data

Using uMLR, the analysis of model simulated data for all 5 soil types identified recommended best sets primarily at depths ≤ 120 cm for typical conditions (Figure 10a, Silt-Best) where the greatest variation in water content occurs (Figure 5a-d). The exception to this general finding is for sand, where the water content variations are deeper (Figure 5e) and the observations extend to 150 cm (Figure 10a, Sand). During wet conditions, the water content varies up to depths of 200 cm for all soils (Figure 4). Therefore, the best sets of observations are deeper and more widely spaced (Figure 10b). Coefficients for the individual soils and conditions are given in Supplemental Information Table S3. Note that the identified RDM best set (Figure 10a,b, RDM) is not best for any particular soil, but represents an acceptable compromise across all of the soil types considered.

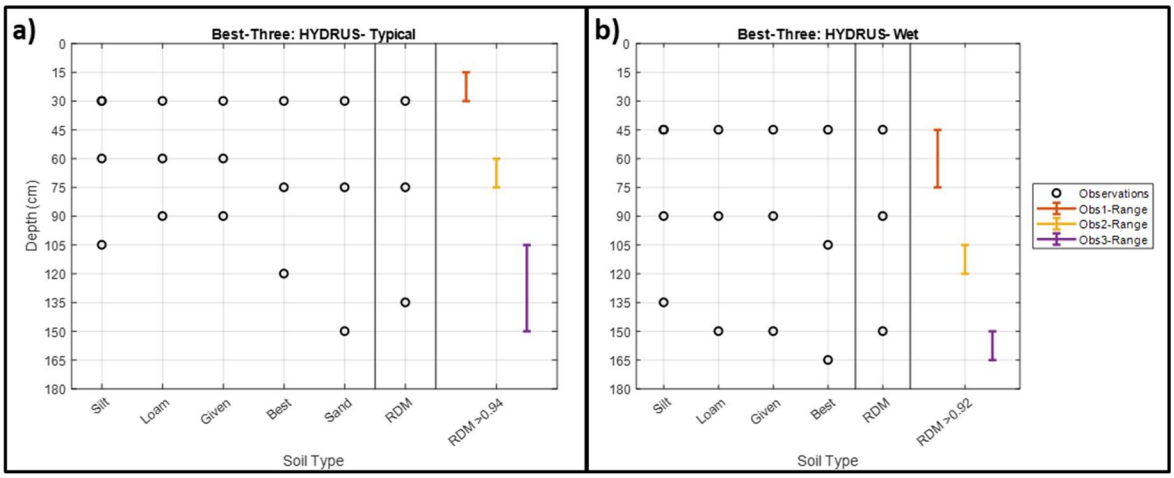

Figure 10: The individual observation sets with the highest R^2 for each site and the best RDM set across all HYDRUS soils for a) typical and b) wet conditions. On the far-right side of the graphs the range of observation depths with an RDM $R^2 > 0.92$ (a) and $R^2 > 0.94$ (b) are shown.

The analysis revealed 75 to 100 combinations of observations that give an $R^2 > 0.92$ across all soil types. For typical conditions, the model-based depth interval recommendations (Figure 10a, vertical bars) are almost identical to the down-sampling recommendations (Figure 9a, vertical bars), except that the model-based ranges have smaller maximum depth for the deepest observation (Figure 9a, 10a, purple vertical bar). For the wet conditions (Figure 10b, vertical bars), the model-based recommendations are considerably deeper and narrower than the down-sampling recommendations (Figure 9b). The RDM set that is best across all precipitation regimes in the HYDRUS model is at 30, 75, and 135 cm with an R^2 of at least 0.90.

Application of the predictive approach

These differences in the model-based and down-sampled observation ranges are to be expected, given that the range of conditions examined with the model is intended to be wider than the actual range of conditions in the field. One strength of the uMLR approach is that it not only provides the correlation to the prediction of interest (R^2) for the single *best set* of three observations, but also for *every possible combination* of three observations (Equation 3). Therefore, we can determine how well the model-based recommendations performed based on the existing field data. First, we determine how well the RDM best set from the simulated data (Figure 10) performed by ranking it among all of the uMLR results

from the down-sampled data (Figure 9). Figure 11 shows the normalized cumulative frequency distribution of the minimum R^2 for the RDM down-sampling sets for the a) typical and b) wet cases; the simulated results are referred to as "HYDRUS Best". The coefficients for the HYDRUS best set, based on down-sampled results for all stations, is included in Table S4 of the Supporting Information.

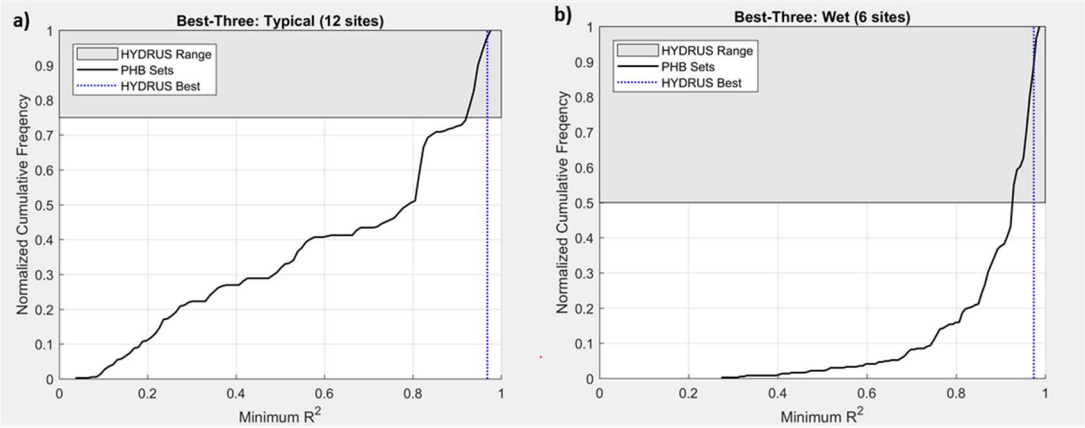

Figure 11: The normalized cumulative frequency of occurrence of minimum R^2 values for the a) typical and b) wet conditions. The RDM best set from the simulated data and its equivalent R^2 during down-sampling is shown in blue. The depth intervals found using HYDRUS (Figure 10) and the number of sets that fulfill this (and their equivalent R^2 values) during down-sampling are shown in gray.

The cumulative frequency curves (black lines, Figure 11) show significantly different minimum R^2 values for wet and typical cases; i.e., a higher proportion of the data sets have high R^2 for the typical conditions. By definition, the RDM set identified by down-sampling is the highest possible minimum R^2. The R^2 for the HYDRUS RDM best set for typical conditions (30, 60, and 105 cm) was 0.97 (blue vertical line, Figure 11a). During wet conditions, the HYDRUS best set (45, 90, and 150 cm) gives an R^2 of 0.97 (blue vertical line, Figure 11b). The relatively small separation of the blue dashed line and the maximum R^2 of the black line gives a visual indication of the excellent performance achieved in this case by use of predictive uMLR-RDM.

We can assess the quality of the recommended depth intervals based on predictive uMLR-RDM. To illustrate this, we examined all combinations of observations sets within the model-based recommended observation ranges and found that they all had R^2 values between 0.92 and 0.97 for typical conditions and between 0.92 and 0.98 for wet conditions (gray shaded regions, Figure 11). For typical conditions, 25% of the total number of possible observation sets fell in this R^2 range (Figure 11a); for wet conditions 50% (Figure 11b). Additionally, the HYDRUS best set across all precipitation conditions (30, 75, and 135 cm) gives a minimum R^2 of 0.93 across all stations. This recommended set gives an R^2 that is lower than the recommended set for each individual condition (wet and typical years), but is still acceptably high across all precipitation conditions.

Conclusion

Limited budgets often restrict monitoring networks. As a result, researchers are required to install fewer sensors than they would like, forcing them to conduct some kind of measurement design optimization effort. These decisions are often made quickly, with little consideration, using rules of thumb or highly simplified assumptions. For large, expensive experiments, computationally intensive, formal

measurement network optimization approaches may be warranted. The universal multiple linear regression (uMLR) method, coupled with robust decision making (RDM), is a simple, computationally inexpensive approach that can be applied to almost any project. We show that uMLR can be used both for the simplification of existing networks (down-sampling), and for network design prior to data collection (predictive).

At the PHB study site, neutron probes were used to monitor store-and-release processes at 156 points throughout the surface barrier (13 depths, 12 locations). The PHB data is used for the purpose of research and is different than practical purposes. A more practical application with the same goal of monitoring water storage in the subsurface through time could have used 25% of the observations (three depths at each location) without sacrificing data worth considerably; a maximum RMSE across all sites of 6.94 mm during the 18-year study period using the predictive RDM best set (Table 4, Supporting Information. For both down-sampling and predictive guidance (before any data were collected), uMLR with RDM was able to demonstrate that only three measurement depths were needed, while identifying the optimal sensor depths. Finally, because uMLR considers all possible observation sets, the results could be expressed as depth intervals within which the three probes could be installed to meet any user-selected performance criteria. Note: there is a risk of sensors failing which could sacrifice data worth when installing less sensors.

The recommendations included in this paper are only applicable for the range of uncertain conditions considered (soil types in Table 1 and precipitation years in Figure 2). But, a similar uMLR-RDM analysis could be performed to consider other sources of uncertainty including changes in barrier thickness; different vegetation; or geostatistical or layered material distributions with depth. The method reported cannot determine the depth of the wetting front, which can be decided based on the water content profile.

Chapter 3

A predictive, model-independent approach to measurement selection to reduce parameter estimation uncertainty

Melissa Clutter, Ty Ferré, and Jeff Klakovich

Target Journal: Water Resources Research

Introduction

Accurate groundwater models require information about the properties and states of the simulated flow system. Unfortunately, constraints on time, financial resources, and accessibility often limit the amount of information that can be collected in the field. Therefore, there are clear incentives to identify data that are most relevant to the modeling objective(s). There are two common approaches for improving model accuracy in the context of model predictions. The first is to reduce prediction uncertainty by identifying parameters that are most important to predictions (Tiedeman et al., 2003). The attributes that are most relevant to predictions are identified using parameter uncertainty analyses (Walker, 1982; Hoybye, 1998) or statistics such as the "value of improved information" (VOII) statistic developed by Tiedeman et al., (2003). Once the attributes are identified, field characterization is focused on identifying data that best constrain these attributes. This provides a direct connection between data collection and model calibration; but, the impact of additional data on the quality of predictions of interest is less direct. The second approach is to identify observations that are most likely to reduce prediction uncertainty. This can be achieved by optimizing network design directly (e.g. James and Gorelick, 1994; Minsker, 2003; Wagner, 1995). Alternatively, prediction optimization can be achieved by first determining the dependence of a prediction on parameters and then assessing the value of observations for constraining estimation of these parameter (Tiedeman et al.,2004).

One method to assess prediction uncertainty is through leverage measures. Leverage is used to identify observations that could potentially have a large impact on predictions prior to model calibrations; the observations that have high leverage or influence are generally very sensitive. Tiedeman et al., (2004) introduced the observation-prediction result (OPR) approach, consistent with leverage, that assesses the impact of prediction uncertainty when observations are added or removed from a field design. This statistical approach serves as a guide for refining monitoring networks, selecting new observation types and locations, prioritizing fieldwork to improve model representation, and understanding the relative importance of observations (Tiedeman et al., 2004). The OPR approach requires data from an *existing set* of observations to calculate prediction uncertainty and to determine which observation locations to *continue* monitoring in the future. Once the sensitives have been calculated there is minimum computational effort and the model does not need to be recalibrated to measure the data. A second approach to network design is jackknifing, which requires omitting one or more observations and then recalibrating a model (e.g. Efron, 1982). Jackknifing can be used to understand observation importance to prediction but can be computationally intensive when there are many observations and/or the model has a long execution time.

Hydraulic conductivity is a controlling parameter for many hydrologic systems; but, it is very difficult to measure directly. As such, it is a common target for inverse analysis. Most data worth analyses require calculation of model sensitivities, which in turn require many forward model runs (Efron, 1982). Some

methods also require existing observations (Tiedeman et al., 2004). For this study, we use a method called universal multiple linear regression (uMLR) as a simple surrogate for a model. This approach is like many machine learning techniques, which seek to find meaningful correlations between targets and observations. But, we use these techniques to examine the relationship between model-predicted observations and model-predicted outcomes of interest as the basis of a data worth analysis.

The uMLR method was initially used to identify a reduced set of parameters that could serve as a substitute more costly colloid mobilizing and transport data collection (Norgaard et al., 2014). Additionally, Clutter and Ferre (in press) formulated the uMLR-RDM approach that uses several predictive model-simulated scenarios to address user-defined uncertainty through robust decision-making (RDM) prior to data collection. All previous studies based on uMLR find the relationship between two system *states* (Norgaard et al., 2014; Clutter and Ferre, in press; Clutter et al., in submission). In this study, we test whether a measure of the information content of different observations sets, based on uMLR-RDM predictive mode (Clutter and Ferre, in press; Clutter and Ferre, in submission), can be used to identify observations that are capable of constraining inversion without the need for a formal sensitivity analysis. This is a new application of uMLR that uses a state to constrain a *property* of the system. This approach, like that shown in Clutter and Ferre (in press) and Clutter et al. (in submission), is predictive. That is, forward model runs are used to train the uMLR analysis without the need for any calibration data. This requires that the model be relatively inexpensive to run so that many model realizations can be used for robust determination of the correlations used for measurement optimization.

Measurement downsampling begins with a large dataset and reduces the number through estimation of measurement uncertainty, directly or indirectly. Alternatively, measurement sets can be constructed beginning with only one observation and adding observations that add value. The objective of uMLR is to build the minimum measurement set by sampling all possible sets of observations. This grid-search approach is fundamentally different than downsampling or sequential construction approaches that only seek the optimal observation set. Specifically, the universal consideration of observation sets provided by uMLR allows for trade-off applications, such as has been shown using RDM (Clutter and Ferre, in press; Clutter et al., in submission), and for greater insight into the underlying reasons for measurement set quality, which is the subject of this investigation.

Methods

Data simulation

We aim to determine if uMLR can inform the selection of observations to constrain numerical inversion and use a simple system for the analysis to draw meaningful conclusions from the results. Specifically, we consider a homogeneous medium with a single layer inclusion with varying K values; the depth and thickness of the inclusion are known. Because both the K of the medium and the K of the inclusion are unknown, we expect to need at least two observations to infer the K values. The objective of this study is to establish how uMLR can be applied for data selection for parameter estimation. There is no reason to expect that the method cannot be extended to more complex problems. In fact, because uMLR was designed to conduct measurement selection at low computational cost, it should become more applicable for more advanced methods and complex systems.

We use MODFLOW2005 (Harbaugh, 2005) to model subsurface in a 100 m soil column airflow due to natural changes in air pressure (Figure 1). Subsurface airflow, can be simulated with Darcy's Law due to low gradients imposed by natural air pressure variation (Massman, 1989). Air flow is simulated as water flow under confined conditions; millibars of air are converted to meters of water before model simulation (1 mbar air = 8.32476×10^{-4} m of water). The USGS Python module FloPy (Bakker et al., 2016) is used to prepare the MODFLOW files for a 1 column x 1 row x 400 layers mesh grid (1 m x 1 m x 0.25 m). In addition, to use MODFLOW the hydraulic conductivity and specific storage are recalculated for air as the fluid.

The soil column consists of three soils with boundaries at 5 and 10 m; the top and middle layers are 5 m thick and the bottom 90 m. The specific storage is set to 2.5×10^{-5} m. The soil configurations are shown in Figure 1. In all cases, the top and bottom layers have the same hydraulic conductivity value (K1) and the middle layer has a different hydraulic conductivity value (K2) (Figure 1a), representing a homogeneous system with a single layer inclusion. Ten hydraulic conductivities were considered (Table 1). We consider two cases: a low permeability inclusion (K1 > K2) (Figure 1b) and a high conductivity inclusion (K2 > K1) (Figure 1c). All possible combinations of the hydraulic conductivity values (Table 1) are considered; 45 combinations where K1 > K2 and 45 different combinations where K2 > K1. We also used 10 homogeneous cases (K1 = K2) for an initial analysis. For clarity, a complete analysis of the homogeneous results is not shown.

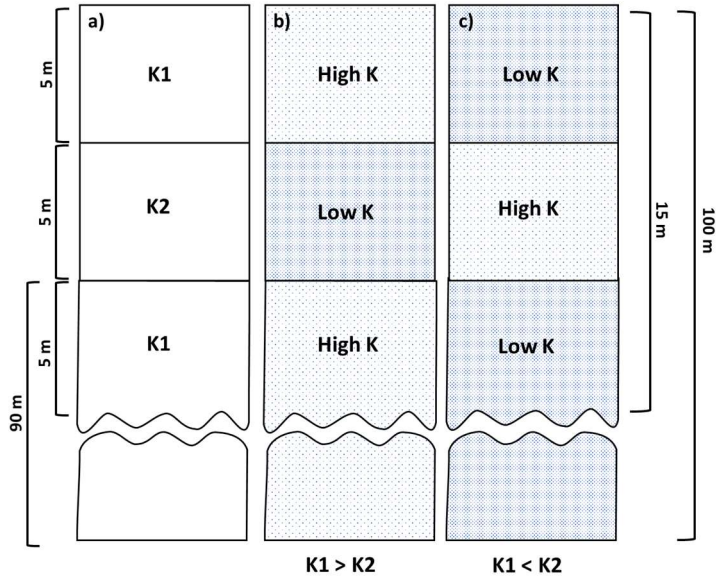

Figure 1: Soil configurations for MODFLOW data simulation in a 100 m soil column. a) The top and bottom soil layers have the same hydraulic conductivity (K1) and the middle layer is different (K2). b) The top and bottom layers have a hydraulic conductivity greater than the middle when K1 > K2. c) The opposite is true when K2 > K1.

Hydraulic Conductivity (m/hr)
1.00E-01
1.00E-02
1.00E-03
1.00E-04
2.15E-02
2.15E-03
2.15E-04
4.64E-02
4.64E-03
4.64E-04

Table 1: The hydraulic conductivity values for the MODFLOW data simulations.

The transient head boundary condition at the surface was the atmospheric pressure measured with a transducer placed one meter above the ground at University of Arizona Tech Park Campus, Tucson, Arizona (Figure 2). Minute-by-minute air pressure data are calculated as the average of six readings per minute from 2018/06/13 00:00:00 to 2018/06/14 23:59:00. The field data were filtered with a high-pass filter removing frequency components with periods longer than 48 hours. The default value for the

convergence criterion was used (1x10^{-5} m for head change). The initial condition is uniform pressure throughout the column equal to the mean of atmospheric pressure (~910 mbar). The model is simulated for 96 hours with one stress period and one time-step per minute; the first 72 hours are used as a burn-in period to overcome the initial conditions and the last 24 hours are used to compute metrics. All boundaries are no flow except the top boundary, which is a specified head boundary (implemented using the time-variant specified-head package).

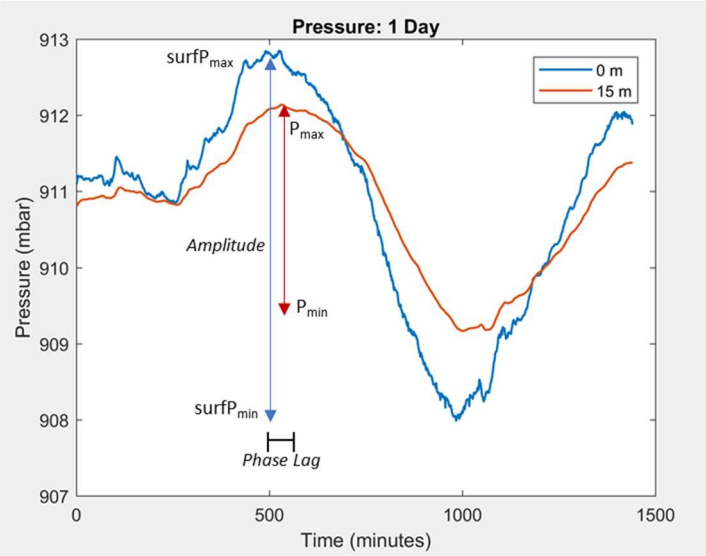

Figure 2: The last 24 hours of atmospheric pressure (0 cm, blue) measured at the University of Arizona Tech Park, Tucson, AZ. The surface data were used as the transient head boundary condition for the MODFLOW2005 data simulation. The MODFLOW-simulated pressure measurement at 15 m depth is shown in red.

Subsurface pressure was not measured for this analysis. Rather, the MODFLOW-simulated pressures at 33 different depths (every 0.5 m from 0 to 15 m with additional measurements at 5,10, and 15 m) were used as virtual data. For testing, corresponding pressure time series were calculated for pairs of K values that were not used for training. An example of the subsurface pressure time series at 15 m depth, corresponding to the atmospheric pressure time series shown, is presented in Figure 2.

As discussed below, numerical inversion used the pressure time series directly. For direct application of uMLR, we needed to calculate a single metric for each test pair of K values that could be regressed against the K values. We examine two features of the time series: attenuation and phase (Figure 2). Attenuation (A) is calculated as the range of amplitude over the 24-hour test period in the subsurface signal compared to the range in the atmospheric signal. Amplitude attenuation (A, mbars) is defined as the peak-to-peak amplitude of the air pressure time series (P, mbars) at each depth subtracted from the amplitude of atmospheric driver (surfP, mbars) (Equation 1; Figure 2).

$$A = (\mathrm{surfP}_{max} - \mathrm{surfP}_{min}) - (\mathrm{P}_{max} - \mathrm{P}_{min}) \qquad (1)$$

The second metric was the phase shift between the subsurface and atmospheric signals (Figure 2). This was determined by time-shifting the atmospheric signal until the correlation of atmospheric to subsurface signal was maximized. Initial investigations suggested that attenuation showed more reliable changes with changes in K values, so this metric was used for our analyses. However, in general, a

composite metric could be developed using multiple time-series metrics. Importantly, any choice of metrics derived from the time series represents some loss of information. This is a key element of this analysis: determining if the simplified uMLR analysis provides reliable information to optimize measurement sets for inversion based on the entire pressure time series.

uMLR

Universal multiple linear regression is a method that assumes that combinations of observations can be linearly related to an outcome of interest (Clutter and Ferre, in press; Clutter et al., in submission). The uMLR approach considers the value of all possible combinations of observations, up to any given subset size, and can be used with pre-existing data to potentially simplify existing network designs (e.g. Norgaard et al., 2014) or predictively prior to sensor installation (Clutter and Ferre, in press; Clutter et al., in submission). In this study, we use model-simulated data to understand the best location to place observations for parameter estimation. We use derived values (i.e. attenuation) for our observations rather than using direct pressure measurements for the uMLR optimization (Figure 3). By using derived values, we transform the pressure data from a time-varying parameter to a single value for regression against a non-varying parameter (hydraulic conductivity) (Equation 1). Therefore, the attenuation observations (at each depth) are a single value that vary with the hydraulic conductivity of the soil. This allows attenuation at different depths to be our candidate observation and the hydraulic conductivity of the soil to be our prediction of interest. If we were considering a homogeneous system, the uMLR analysis would be applied as described in Clutter and Ferre (in press).

$$K_j = a_0 + \sum_{i=1}^{n} a_i A_i \qquad (2)$$

where a series of attenuation (A) values at the chosen observation depths, *i*, is linearly related to a series of hydraulic conductivity values (K_j), a_0 is the y-intercept, a_i are the regression coefficients, and n is the number of observation depths.

For heterogeneous conditions, there are multiple dependent variables and the attenuation observations are correlated to K pairs (K1, K2). To find the correlation between the K pairs and attenuation we must use *multivariate* (multiple) linear regression (Equation 3).

$$K_{jk} = a_{0k} + \sum_{i=1}^{n} a_{ik} A_{ij} \qquad (3)$$

where K_{jk} is the hydraulic conductivity for the j^{th} set of K pairs, with k representing the upper/lower or middle layer, a_{0k} is the y-intercept for the corresponding layer, n is the number of observation depths, a_{ik} is the regression coefficient for the i^{th} observation depth for the corresponding layer, and A_{ij} is the attenuation for the i-th observation for the j^{th} K pair. The A_{ij} values can represent any transformation applied to the calculated attenuation values (see below for further detail on transformation). Note: uMLR does not explicitly consider parameter correlation. One of the objectives of this investigation was to determine if this simplification negatively impacts the utility of uMLR for identifying optimal observation sets.

Multivariate uMLR uses RMSE (rather than R^2) as the coefficient of determination to determine the information content of the observation set. RMSE values are calculated (Equation 4) independently for K1 and K2. Therefore, some observations can be informative for predicting K1 but not K2, or vice versa. Alternatively, uMLR selection can be based on a combined measure of RMSE considering both K estimates:

$$RMSE = \sqrt{\frac{\sum_{n=1}^{numK} \left(\frac{KuMLR_n - Kactual_n}{Krange}\right)^2}{numK}} \quad (4)$$

where KuMLR [m/hr] is the K predicted by the uMLR linear regression, Kactual [m/hr] is the MODFLOW-model simulated K, K range [m/hr] is the range of actual K values, and numK [-] is the number of K values considered. The uMLR analysis identifies the most informative sets as those with the lowest RMSE (in contrast to previous applications, which sought to maximize R^2).

For this study, the uMLR analysis requires a forward model run for each pair of K values. But, every combination of observations can be analyzed using multiple linear regression without having to rerun the model. For example, if 10 K values are considered and there are 100 potential combinations, uMLR requires only 10 forward model runs for all 100 regressions (Equation 2; Equation 3). In contrast, to assess the value using inversion we must do tens to hundreds (or more) forward model runs for each K pair and each measurement set considered. If we assume the inversion prediction requires 100 forward runs, the same number of observation combinations would require ~100,000 forward model runs. As a result, uMLR allows for a much broader consideration of potential measurement sets than methods based on full inversion.

Inversion

For inversion, we use simulated data in a model-independent parameter estimation and uncertainty analysis tool (PEST) (Doherty, 2016) to infer hydraulic conductivities. Specifically, we use the same conditions as presented above to forward model subsurface pressure observations in MODFLOW but restrict the simulation to a limited number of observation depths and one K pair (K1, K2). The observation depths are determined by the RMSE from the uMLR results and the two hydraulic conductivities considered (1.93×10^{-3} and 1.593×10^{-2} m/hr) were not included in the training set (Table 1). For K1 > K2, K1 is 1.93×10^{-3} m/hr and K2 is 1.593×10^{-2} m/hr. The configuration is opposite for the K2 > K1 case. Note: we do not use attenuation values for the inversion, but rather use the MODFLOW-generated pressure time series at each specified depth calculated for the 'truth' conditions.

PEST is a model-independent numerical calibration tool that back-calculates and matches a model's outputs to model inputs (e.g. parameters) (Doherty, 2016). PEST uses the information from these calculations to write and replace information in the MODFLOW Layer Property Flow (LPF) data file. The LPF file is manipulated and MODFLOW is rerun (with new parameters) until the convergence criteria specified in PEST are met. The metric PEST aims to match for time-varying pressure is the sum squared error (SSE) for all time steps (Equation 5).

$$SSE = \sum_{i=tmin}^{tmax} (Pinital_i - Pnewguess_i)^2 \quad (5a)$$

Where P_{inital} is the initial pressure values (or the P from the previous iteration), $P_{newguess}$ is the pressure values from the current iteration, t_{min} is the start time, t_{max} is the end time, and the time step is determined by the data.

$$SSE_{signal} = \sum_{i=tmin}^{tmax} (Psurf_i - P_i)^2 \quad (5b)$$

where P is the pressure signal at depth and Psurf is the atmospheric signal.

To reiterate a difference between uMLR and inversion, SSE_{signal} is influenced by all differences between observed and simulated signals (e.g. including attenuation, phase shift, and all other sources of difference), whereas uMLR uses only one element of the difference between the observed and simulated time series (attenuation).

By default, PEST will cease execution once ONE of the following criteria are met 30 iterations or the variable PHIREDSTEP is less than 0.005 for four consecutive iterations.

$$PHIRESTEP = \frac{\varphi i - \varphi min}{\varphi \mathrm{i}} \quad (6)$$

where φi is the objective function for the iteration and φmin is the minimum value of the objective function for all iterations. The initial K during parameter estimation is 1 x 10^{-6} m/hr; the K is log-adjusted with bounds 1 x 10^{-6} to 1 x 10^{5} with no regularization.

Once the model converges, PEST provides inferred K1 and K2 values (Kinv). These values are compared to the forward model K values (Kactual) to find the percent error for the inversion. The percent error from the inversion represents the inversion success (Equation 7), with a low percent error representing a more successful inversion.

$$\%\, Error = \frac{|Kinv - Kactual|}{Kactual} * 100 \quad (7)$$

We compare the accuracy of the inversion, for the restricted observation sets, to the predicted quality (RMSE) (Equation 4) from uMLR (for the same set) to understand if we can use uMLR to choose observations to constrain the inversion to accurately recover the mean K pair. While there is no reason to expect that the uMLR-based RMSE will be directly related to the inversion error, for uMLR to be useful for identifying observation sets to constrain inversion, the relative values of both metrics should be comparable for both metrics. That is, observation sets are chosen because they have a low RMSE in the uMLR analysis; this should indicate low inversion error for the recommendations to be reliable.

Results and Discussion

MODFLOW simulated data

Subsurface pressure measurements were simulated in MODFLOW and the attenuation was calculated at the 33 specified depths using Equation 1. Figure 3 shows the attenuation with depth for the two soil configurations (high or low permeability inclusion); the horizontal dashed lines represent the layer boundaries. Note: the domain extends to 100 m, but only the top 15 m is shown as this is the region that includes candidate observations. For both configurations, all K pairs (45 configurations) are shown as individual series. The color of the lines refers to the value of K1. The attenuation is zero at the surface and increases with depth and the change in attenuation with depth is greater through the low K layers than through the high K layers.

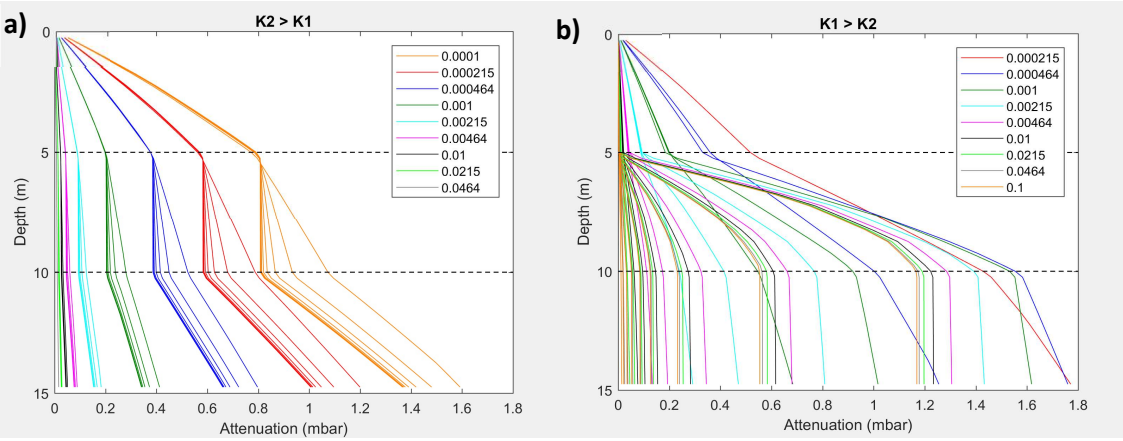

Figure 3: Attenuation versus depth for a) K2 > K1 and b) K1 > K2. All 45 different layer configurations are shown for each case and the colors represents the 33 different depths. The layer boundaries are located at 5 m and 10 m depth (black dashed lines).

For the high permeability inclusion (K2 > K1) (Figure 3a) there is an increase in attenuation with depth through the upper layer; the smaller the value of K1, the greater the change in attenuation with depth. Additionally, the change in attenuation with depth from 0-5 m is not impacted by the K value of the inclusion. Once the signal crosses 5 m, the attenuation with depth is determined by the value of K2; attenuation is higher for low K2 values. The attenuation with depth through the bottom layer (> 10 m) is equal to that in the top layer (5 m) for each K1 value. In Figure 3a, the grouping of colors indicates that the K pairs with the same K1 values, have similar amplitude values with depth.

For the low K inclusion case (K1 > K2) (Figure 3b), the low K inclusion has some impact on the attenuation with depth in the top layer for the three lowest K1 values (2.15e-3, 4.64e-3, and 1e-3); when the inclusion has a lower K value, the gradient through the upper layer is reduced. Once the signal crosses the 5 m boundary (K2), there is a large increase in attenuation with depth, which depends on the K2 value; the K1 has an influence on the magnitude of this attenuation. For most cases there is very little additional attenuation for depths > 10m; most of the energy was lost through the low K inclusion and the signal is damped below 10 m.

We expect that the attenuation observations will be the least informative if 1) there is little variation in attenuation with depth or 2) the signal is already highly damped when it reaches the observation depth. For example, in the case of the high inclusion layer (Figure 3a) there are only small changes in attenuation with depth through the high K inclusion (5-10m) for most K pairs. For most K pairs, we might expect that the attenuation observations between 5 and 10 m would be redundant. As a result, it would not be useful to have more than one observation in the inclusion. Similarly, most observations below 10 m for the low inclusion case (K1 > K2) show little attenuation with depth (Figure 3b), indicating that more than one observation in the lower layer would not be informative. These general conclusions are not particularly helpful, however, given that it is unlikely that choosing two observations in the same layer would be a natural design choice.

Simple uMLR with transformation

We began our analysis by performing traditional uMLR (Clutter and Ferre, in press) (Equation 2) for the simplest homogenous case. That is, using attenuation as the candidate observation, the K of 10

homogeneous soils as the prediction of interest (K values given in Table 1), and R^2 as our coefficient of determination.

The best-three observations (5, 12.25, 14.25 m) yielded a low R^2 (R^2 = 0.3298) for such a simple soil configuration (Figure 4a). We attributed the poor fit (low R^2) to the nonlinear relationship between attenuation and hydraulic conductivity (Figure 4b). To improve the uMLR prediction we explored transformations of the derived observation (attenuation) to linearize the relationship between the outcome of interest and the candidate observations. Note that it is also possible to transform the prediction of interest to linearize the problem, but this precludes applying different linearization approaches to different candidate observation types.

We tested the linearity of several transformations of the attenuation against K by conducting a standard regression and constructing a residual plot, including: a logarithmic model (K α log(Attenuation)), reciprocal model (K α 1/Attenuation), and an exponential model (K α $10^{Attenuation}$). The best transformation method (R^2 = 0.9995) for the homogeneous case (Figure 4c), was a reciprocal model (Figure 4d).

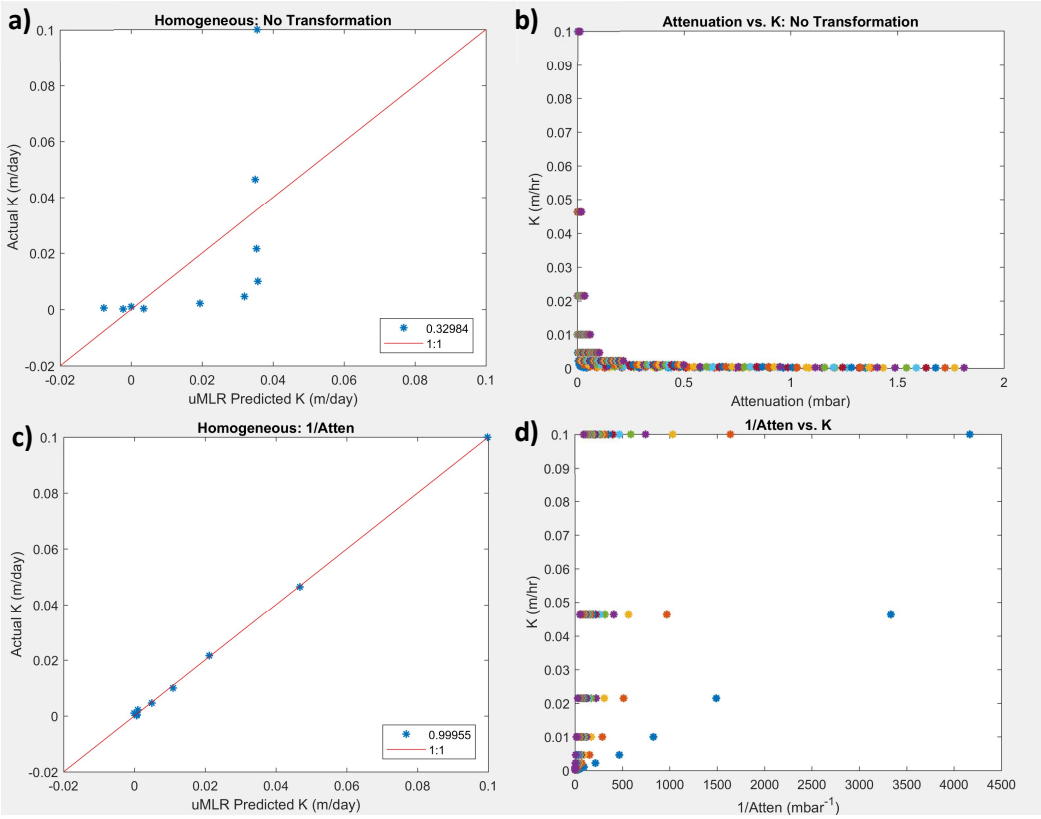

Figure 4: Multiple linear regression for the homoegenous case (K1 = K2) based on the uMLR best 3 (5, 12.75, 14,25 m) using a) original attenuation values and c) transformed attenuation values (1/attenuation). The b) original relationship between K and attenuation and d) transformed relationship are shown. The different colors represent the 33 observation depths.

Multivariate uMLR suggestions

Prior to inversion, we used the transformed attenuation variables (Figure 4d) to do a multivariate uMLR analysis (Equation 3) for the heterogeneous cases (Figure 1b,c). The RMSE values, for all combinations of

1-3 observations, were calculated for K1 and K2, separately (Equation 4). Our working hypothesis was that the relative magnitude of the RMSE values represents the value of information of each observation set to constrain inversions. A small spread of RMSE values, for any set size, suggests that there is little value in carefully selecting observation placements (i.e. combinations of observation depths perform similarly).

Figure 5 shows the RMSE values for every possible set comprised of a given number of observations for the high K inclusion case (K2 > K1). In this case, uniformly low RMSE suggests almost any set could potentially predict the top/bottom layer (Figure 5a) and uniformly high RMSE values suggest there are likely no sets that could accurately predict the K of the inclusion (Figure 5b). Additionally, the minimal improvement achieved in the lowest RMSE by considering more than little is to be gained by adding a third observation.

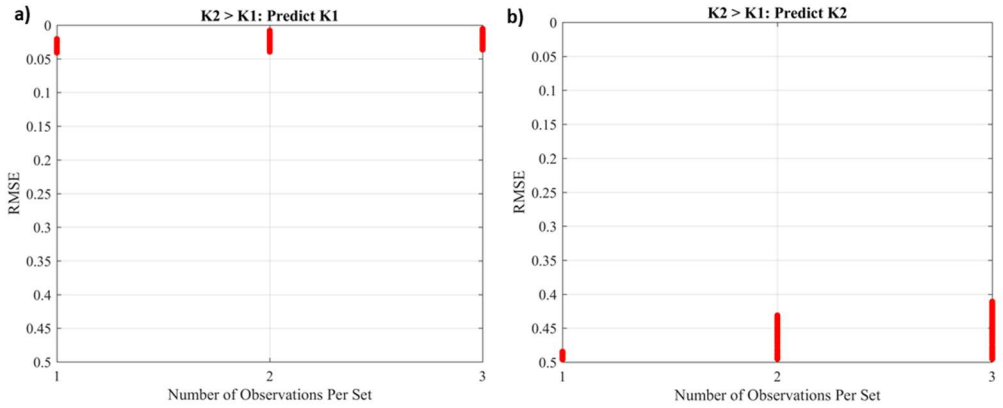

Figure 5: The uMLR results for all combinations of sets of 1-3 observations for predicting a) K1 and b) K2 when K2 > K1.

Figure 6 shows the RMSE values for each set for the low K inclusion case (K1 > K2). In general, there is a larger spread of RMSE values among sets than for the high inclusion case, particularly when predicting K1 (Figure 6a). The larger spread of RMSE values suggests that there is greater value in carefully selecting observation depths and the choice of depths has a strong influence on the information content of the observations. For predicting K2, all sets (of all sizes) have an RMSE < 0.2 (Figure 6b) and any combination of observations could potentially predict K2 (although some sets performing better than others). The RMSE values range from 0.17-0.47 for predicting K1. To summarize, uMLR suggests that it is likely to be more successful to infer the permeability of the background and the inclusion for a high permeability inclusion in a lower K background than for a low permeability inclusion in a higher K background. Further, that there is greater value in optimizing the measurement depths for a low K inclusion. Finally, while there is some level of user-preference in the decision of acceptable improvement with added observations, we decided to limit further analyses to sets of two pressure observations.

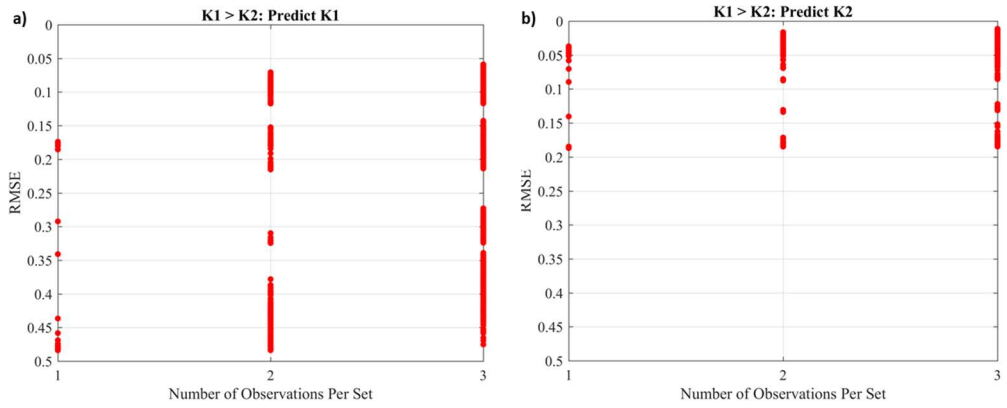

Figure 6: The uMLR results for all combinations of sets of 1-3 observations for predicting a) K1 and b) K2 when K1 > K2.

Multivariate inversions based on uMLR results

The goal of the inversion analysis is to answer the question, *is the ranking of observation sets by uMLR useful for predicting the relative inversion performance*? Therefore, we compare the ranking of the uMLR results (Figure 5,6) (Equation 4) against the inversion accuracy of PEST (percent error) (Equation 7).

K2 > K1

Figure 7 shows a comparison of the inversion accuracy and ranking of uMLR best sets for the high permeability inclusion (K2 > K1) case. As expected, the sets with one observation had the highest RMSE and percent error (Figure 7, orange dots) and sets with two (or more) observations had lower RMSE and percent error (Figure 7, blue and gray dots). The results also confirm our choice to focus on sets of one or two observations.

For predicting K1, the uMLR ranking suggests that all sets have high informational value (RMSE < 0.2) (Figure 5a). This is confirmed by low percent error (<= 10%) for all inversion results, of all set sizes (Figure 7a). There is still a correlation between the RMSE values and percent error for predicting K2, (Figure 8b). However, it appears that the RMSE values are not transferrable between the two cases. For example, for predicting K1, the RMSE only ranged from 0.01 to 0.04, with all inversions having an error less than 10%. In contrast, for predicting K2, the uMLR ranged from 0.44 to 0.50. This is a similar range in RMSE, with a very different mean value. Furthermore, there were cases for which the percent error for inferring K2 was less than 10% with RMSE values that are much higher than for inferring K1. In summary, it appears that uMLR (using attenuation) can provide relative ranking information for inversion accuracy. However, it is unclear whether the magnitude of the RMSE value can be transferrable among cases or for different inversion targets.

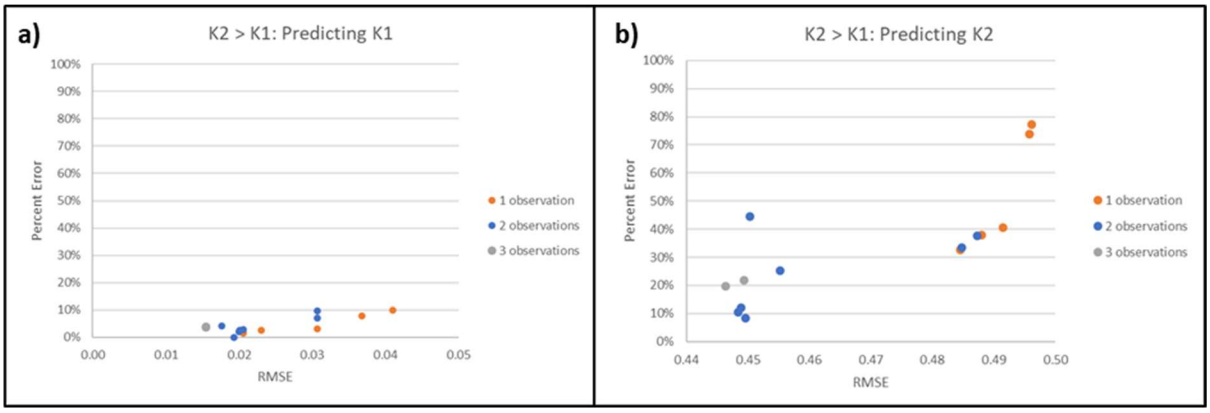

Figure 7: The relationship between percent error from the PEST inversion and the RMSE when predicting a) K1 and b) K2 for the K2 > K1 case. Observation sets with 1 observation are shown in orange, two observation in blue, and three observations in gray.

Figures 8 a and b show the depth of observations for sets of 1 observation. The black dots show the uMLR RMSE for the corresponding depth; the red diamonds indicate the magnitude of the inversion error at that depth. (Note: all markers used to represent inversion accuracy for the remainder of the paper have identical multipliers and can be compared across panels.) For predicting K1, uMLR suggests that the single best observations are in the top layer and the worst are in the bottom layer and the opposite is suggested for predicting K2 (Figure 8b, black dots). The inversion accuracy follows the trend of RMSE; sets with high RMSE had the highest error (large symbols) and sets with the lowest RMSE have the lowest error.

Sets of two observations have similar trends to those with one observation. The bottom panel of Figure 8 (c and d) show the RMSE for all combinations of two observations (with no repetitions). The x and y-axes are the depths for each observation and the contour lines are the RMSE for each set; half of the plot is filled with NaN values to exclude repeated set combinations (observation 2 is always deeper than observation 1). The soil layer boundaries are shown as white dashed lines. The sets tested through inversion are shown as points on the contour plot (white diamonds) and the size of the marker indicates the percent error for each prediction (large symbols represent higher inversion error and the symbols are scaled identically across all color fill plots). The best sets of two observation for predicting K1 have the shallow observation in the top layer (< 5m) and the worst are when both are in the lowest layer (> 10m). Sets with observations in the inclusion or near the bottom of the top layer have intermediate RMSE values. These results are confirmed by inversion, as shown by sets with low inversion error (small white diamonds) coinciding with areas of low RMSE and sets with larger inversion error (larger white diamonds) corresponding with higher RMSE values. For most cases, the percent error is higher for predicting K2 than K1.

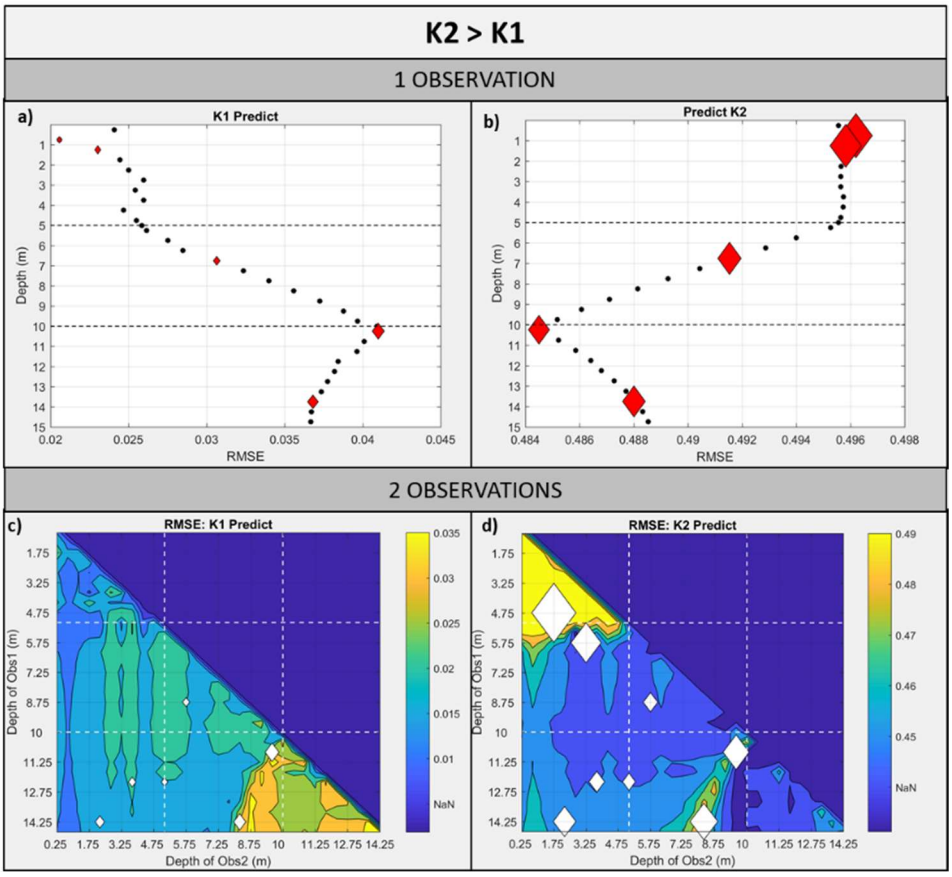

Figure 8: All combinations of one observation for predicting a) K1 and b) K2 are shown as black dots. The sets that were chosen for inversion are shown as red diamonds and the relative inversion accuracy is demonstrated by the size of the marker. The layer boundaries are shown by black dashed lines. The RMSE for all combinations of two observations for predicting c) K1 and d) K2 are shown in a contour plot; the colorbar shows the range of RMSE values. The sets used for inversion are shown in white and the relative inversion accuracy is demonstrated by the size of the marker. The layer boundaries are shown as white dashed lines. Note: all markers used to represent inversion accuracy (diamonds) have identical multipliers and can be compared across panels.

The RMSE values given by uMLR for each observation set (Figure 8) are the mean error over all K pairs examined (Table 1) (Equation 4). In contrast, the inversion performance is calculated for a specific realization of the inversion K values. As a result, uMLR may indicate that an observation set is likely to be informative (on average), but that set may perform poorly for a specific set of K values. To understand if the RMSE given by uMLR for all combinations of K pairs (based on values in Table 1) is an appropriate representation of the quality of the single K pair used for inversion (1.593×10^{-2} m/hr and K2 = 1.93×10^{-3} m/hr) we must understand the distribution of the uMLR error for the range of K values considered (Table 1). The uMLR error for a single observation at 13.75 m depth for all K pairs considered is shown in Figure 9. The error for each pair is contoured, the values of the pairs are plotted as red circles, and the values of the single pair of K values used for inversion are shown as a white diamond.

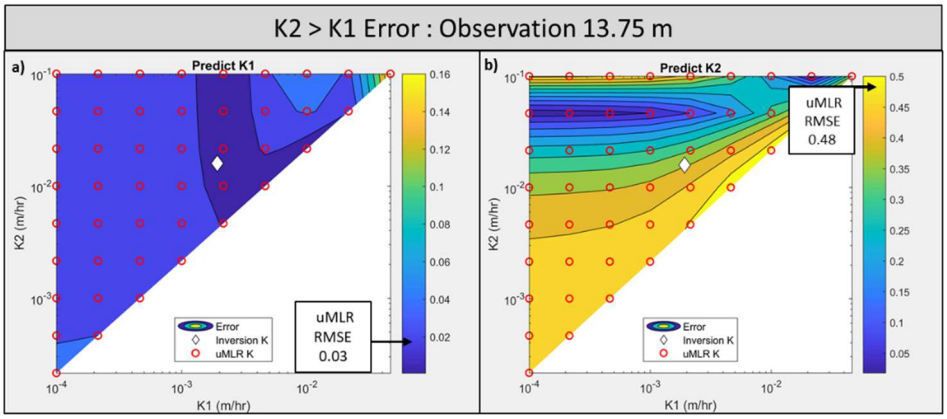

Figure 9: The uMLR error for predicting a) K1 and b) K2 for the K2 > K1 case. All K pairs considered for uMLR are shown by red circles. The RMSE (0.03 for K1 and 0.48 for K2) is a mean of all uMLR error values. The white diamond is the K pair considered for the inversion (K1 = 1.93E-3, K2 = 1.593E-2).

For predicting K1 (Figure 9a), the uMLR method gives a low error for almost all K pairs considered (red circles). Because most K pairs have a similar error, the RMSE ($RMSE^2$ = 0.0012) is a robust representation of the K pair used for the inversion (white diamond) and any other K combination considered in this range. Additionally, the low RMSE when predicting K1 agrees with the low percent error from the inversion (percent error = 7.73%). For predicting K2 (Figure 9b), the uMLR method gives a range of errors depending upon the K combination considered (red circles). Unlike the K1 prediction, which had a low error for all K pairs (Figure 9a), many of the K pairs considered by uMLR (red dots) have a relatively high error for predicting K2 (Figure 9b). Of course, when selecting a measurement set we don't know which K values will exist; that is the objective of the monitoring effort. So, some average error must be used. But, for this case, the RMSE provided for predicting K2 could be overly optimistic if the actual K pair corresponded with a yellow area, or overly pessimistic if the actual K pair was in a dark blue area.

Multivariate linear regression (Equation 3) fits K1 and K2 simultaneously and the fit of one K value could be comprised at the expense of predicting the other. However, there is some disconnect between using uMLR to find each K value (K1, K2) individually and traditional inversion, which by its nature seeks to fit both. There is potential to use uMLR-RDM (Clutter and Ferre, in press; Clutter et al., in submission) to seek a tradeoff for the identifiability of K1 and K2. For example, if K1 were a higher priority than K2, the uMLR-RDM approach could be structured to identify sets that are in the top 5% for identifying K1 while also being in the top 30% for identifying K2. This application will require additional analysis, but it holds the promise of allowing a user to choose weights to place on identifying each K value. Standard inversion methods allow the user to place weights on observations. But the uMLR-RDM method could further personalize inversion by allowing the user to emphasize the accuracy of inversion of specific parameters.

K1 > K2

For the low permeability inclusion case (K1 > K2), uMLR suggested low RMSE values (< 0.20), for predicting K2 for all sets considered (Figure 10b). The percent error from inversion for predicting K2 was relatively low with a slight increase in error for RMSE values closer to 0.2 (Figure 10b). For predicting K1, uMLR suggested a range of potential inversion accuracies (0.07 > RMSE > 0.49) (Figure 6a) depending on the depths of the observations. The PEST inversions gave a range of percent error; however, the trend was not what we expected. In most cases, when inferring K1, low RMSE values were associated with

high percent error and high RMSE values had a low percent error (Figure 10a). That is, uMLR was not informative regarding the optimal observations for inversion.

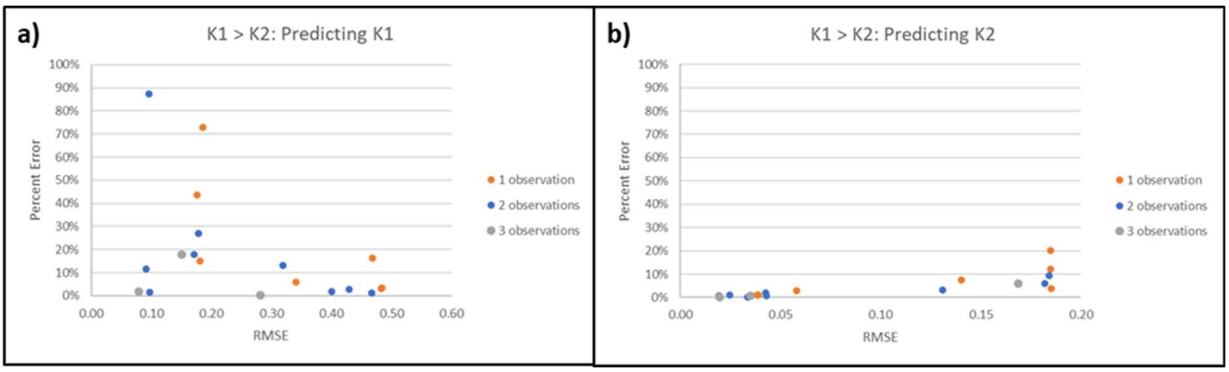

Figure 10: The relationship between percent error from the PEST inversion and the RMSE when predicting a) K1 and b) K2 for the K1 > K2 case. Observation sets with 1 observation are shown in orange, two observation in blue, and three observations in gray.

We can look at the depths for each case to gain insight into the relationship between RMSE and inversion error. As seen in Figure 10b, when predicting K2, the relationship between RMSE and inversion error is what we expect. For one observation, where the RMSE is high at shallow depths (Figure 11b, black dots), the inversion error is also relatively high (Figure 11b, larger red diamonds). Similarly, the RMSE and inversion error are both low at greater depths. For predicting K2 with two observations, sets with two observations in the shallow subsurface (< 5 m) have the highest RMSE (Figure 11d and inversion error (Figure 11d, larger white diamonds). All other sets of two observations have a low inversion error (Figure 11d, larger white diamonds) and RMSE (Figure 11d, blue contours).

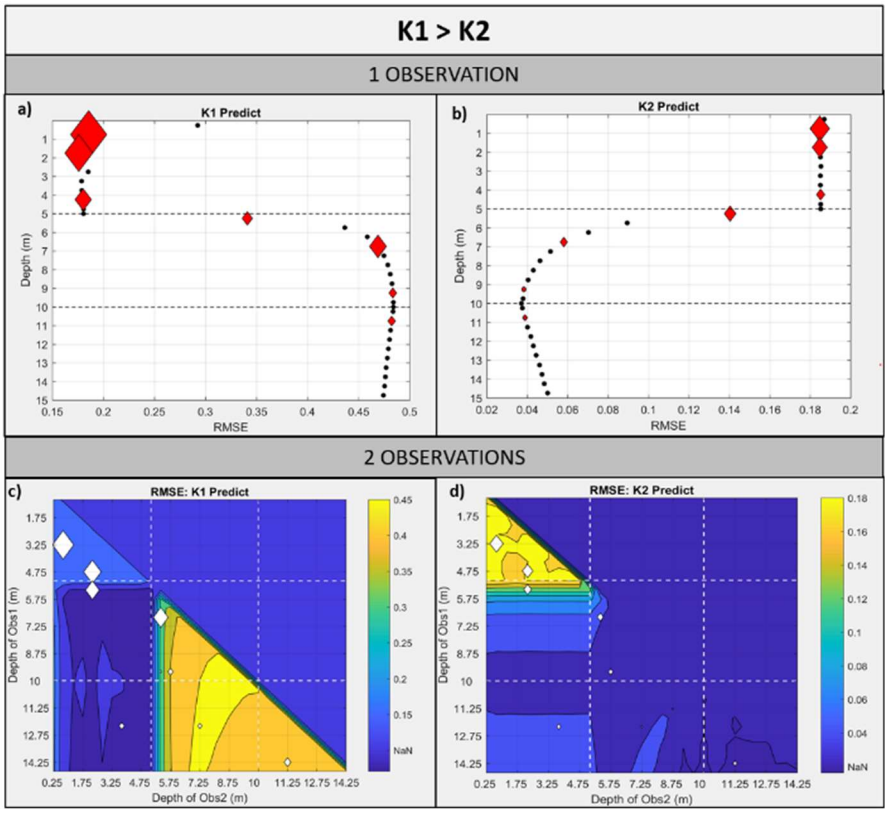

Figure 11: All combinations of one observation for predicting a) K1 and b) K2 are shown in black. The sets that were chosen for inversion are shown in red and the relative inversion accuracy is demonstrated by the size of the marker. All combinations of two observations for predicting c) K1 and d) K2 are shown in a contour plot. The colorbar gives the RMSE values. The sets used for inversion are shown in white and the relative inversion accuracy is demonstrated by the size of the marker.

For predicting K1, the trend does not conform to our expectations, particularly for shallow (< 5m) and deep (> 10m) observations (Figure 10a; Figure 11a,b). For sets of 1 observation (Figure 11a), the inversion error is lowest in the deep subsurface (smaller red diamonds) where the RMSE is highest (black dots). Therefore, the uMLR approach recommends deep observations to give less accurate prediction of K1 (high RMSE), but the actual inversion accuracy is high (low percent error) from the PEST inversion. Two observations in the shallow subsurface (< 5m) have low uMLR RMSE (Figure 11c, blue contours), but the inversion accuracy is also low (Figure 11c, larger white diamonds). When both observations are at deeper depths (> 5m) the RMSE is higher (Figure 11c, yellow contour), but the inversion accuracy is also higher (smaller white diamonds).

The contrasting results for predicting K1 for the low K inclusion case require explanation before uMLR can be used to guide data collection to constrain inversions. The differences likely lie in the different bases for assessing goodness of fit for uMLR and inversion. Specifically, RMSE (Equation 4) for each depth is determined by the fit of the 1/Attenuation and K values (Figure 4c). In contrast, the prediction for each inversion is based on fitting the pressure time series. In this context, the inability of uMLR to predict the inversion quality of some observation sets, particularly for the low inclusion condition, may be attributed to loss of information in reducing the pressure time series to a single value of attenuation. However, this does not explain the high correlation between 1/attenuation and K for shallow

observations, which is not shown for inversion based on shallow observations. In fact, this suggests that standard inversion approaches, based on fitting predicted time series, may be improved by reducing these time series to simple metrics.The difference in the metrics used for uMLR and inversion can be examined visually (Figure 12). Here, we calculated the SSE_{signal} for each depth (Equation 5b), for the same time period as the attenuation metric calculations (Equation 4). The SSE for each depth is shown, with the color of the lines indicating the K1 value.

For the high inclusion case (K2 > K1), we see very similar trends in SSE (Figure 12a) and attenuation (Figure 12b). The SSE increases with depth through the low K top layer (> 5m), has less variation through the high K middle layer, and increases with depth again through the low K bottom layer (> 10m) (Figure 12a). The magnitude of both metrics at depth is primarily determined by K1 and are clustered by K1 in the plot. For example, metrics with the lowest K1 (0.0001 m/hr, orange) will always have the highest attenuation and SSE, no matter the value of K2. Attenuation is a good representation of how the signal changes with depth. As a result, the RMSE values calculated using attenuation are representative of the inversion quality (Figure 8).

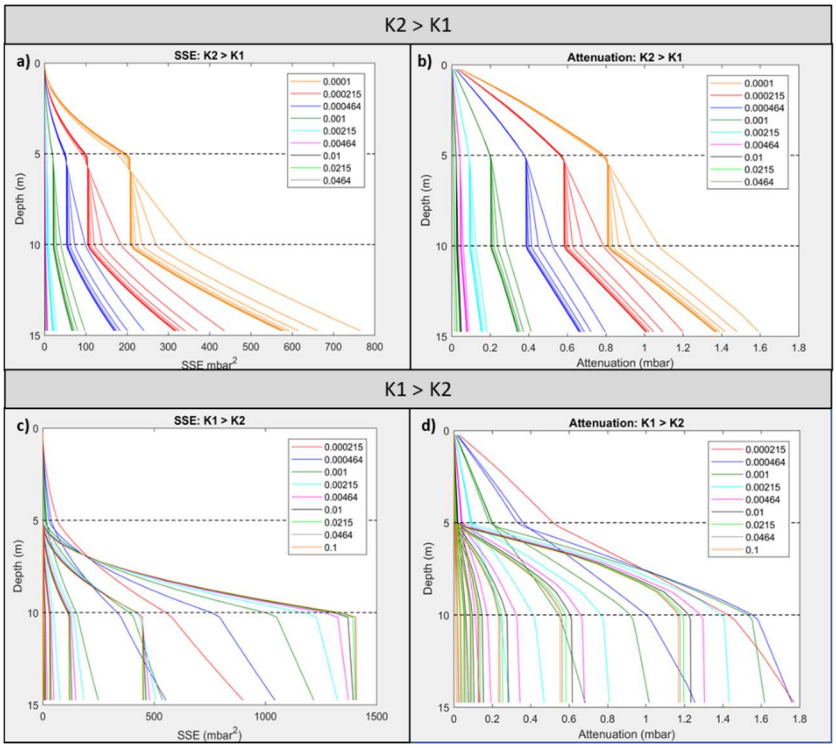

Figure 12: The SSE (Equation 7) and attenuation (Equation 4) by depth for all K pairs for the high K inclusion (a,b) and low K inclusion (c,d) cases. The colors of the lines are determined by the value of the K1.

For the K1 > K2 case, the SSE (Figure 12c) and attenuation (Figure 12d) vary differently with depth. There is less variation in SSE (Figure 12c) with depth through the top layer (< 5m) than for attenuation (Figure 12d). Both metrics vary similarly through the middle layer. And for some K pairs there is more variation in SSE at depth (< 10m) than for attenuation. For deeper observations, there may be more information provided by phase lag, so uMLR would was not conditioned on the most informative metric at this depth; because of this loss of information, uMLR was not able to identify informative deep data sets for inversion. In contrast, the use of attenuation to condition uMLR appears to provide more useful information for inversion than using the entire pressure time series. For both metrics, the variations are

primarily determined by the low K inclusion (K2), and the plots are not clustered by color as in the high K inclusion case (Figure 12a,b) where the magnitude of the variations were primarily determined by the low K background layer.

Conclusions

We often use field observations in hydrologic models to infer difficult hydraulic properties to measure, such as K. The parameter uncertainty from a parameter estimation model is dependent on the value of information provided by the field observations. One approach to constraining the observations is to use models; unfortunately, most models require pre-existing observation data and/or are computationally intensive. In this paper we use uMLR, in predictive mode, as a simple surrogate for a model to constrain observations to increase parameter estimation accuracy in PEST. By using the uMLR recommended best depths for observations, in many cases the inversion error was low. This suggests that a simple tool like uMLR could be used to improve inversion accuracy. In other cases, the average ranking provided by uMLR over a range of K values was too conservative or too optimistic for the accuracy of the inversion. It is possible that these discrepancies are due to changes in the information content of the pressure time series when it was converted to a metric to use for uMLR. Further examination of this will improve the use of uMLR for data selection and may improve hydrologic inversion more generally.

Chapter 4

Research Summary and Contribution

Budget limitations often restrict the type and number of sensors that are installed at a field site. Universal multiple linear regression is a simple, robust method for optimizing a network design. Norgaard et al., (2014) first introduced uMLR to identify a reduced set of parameters that could be measured as a surrogate for performing more costly analyses. Specifically, they examined the ability to use pre-existing highly disparate soil analyses including soil electrical conductivity, water content at specific pressures, and soil particle size fractions as indicators for colloid mobilization and transport. My work expands on the uMLR approach introduced by Norgaard et al., (2014) by presenting the method in the context of new field scenarios, combining uMLR with RDM, exploring uMLR as a predictive approach, and by optimizing measurement networks specifically for parameter estimation.

I extend the uMLR approach to three new applications. First, I use water content and temperature observations to predict time-varying total water storage in an ephemeral stream channel (Paper 1). Next, I use water content observations to predict time-varying total water storage in an evapotranspiration surface barrier (Paper 2). Lastly, I use uMLR to identify the best subsurface pressure observations for predicting hydraulic conductivity (Paper 3).

In addition to extending uMLR to new applications, I merged it with Robust Decision-Making (RDM). RDM was a concept introduced by Lempert et al., 2013 for high level policy decisions (Groves et al., 2008, Hine and Hall, 2010, Lempert and Groves, 2010). The basic concept of RDM is to find the best design, that although suboptimal for any one scenario, is the best across many plausible scenarios. Typically, optimization approaches are designed to find a single best design. However, the goal of uMLR is find the relationship between *all possible combinations* of observations to the prediction of interest. This unique feature of uMLR provides the opportunity for additional analyses such as uMLR-RDM. The uMLR-RDM method uses all observation combinations to find an observation design that is robust under uncertainty and would perform satisfactorily over an ensemble of scenarios; a uMLR-RDM best observation set can cope with a range of user-defined errors. Additionally, the uMLR-RDM can define acceptable *depth intervals* for sensor installation (Paper 2).

Downsampling of existing data is useful for simplifying network designs and to understand the most important measurements for calculating the outcome of interest (e.g. Norgaard et al., 2014; Tiedeman et al., 2004). Downsampling is a relatively straightforward processes and most statistical data worth approaches manipulate existing field designs to understand observation sensitivity to prediction uncertainty. Unfortunately, many field experiments do not have any pre-existing data. Therefore, the predictive approach to uMLR is a valuable tool for determining the best locations for sensor installation prior to data collection. The predictive tool was proposed in Paper 1 and tested in Paper 2 with the field data from the Prototype Hanford Barrier site.

In hydrology we often use data to constrain models for the estimation of hydraulic properties. However, there is a need to constrain the input data, prior to modeling, for a reduction in parameter estimation uncertainty. Specifically, in Paper 3, we use uMLR to reduce predictive uncertainty by improving the value of the observations. Previously, researchers have improved the value of observations either through optimization using pre-existing data (leverage) or through computationally intensive optimization approaches that require model calibration for each potential observation combination

(jackknifing). The two strengths of using uMLR to study the value of observational data is 1) the predictive approach allows users to optimize designs prior to data collection and 2) uMLR is a model-independent approach that is computationally inexpensive.

Generally, uMLR is a unique and useful approach to data worth analyses because it is very simple and, therefore, computationally inexpensive; it is model-independent and uses a linear relationship. As a result, a researcher can afford to conduct a global search of possible measurement combinations. This can uncover not only the best measurement set, but it can also provide insight into the potential value of expending effort to optimize data collection. Furthermore, because uMLR is relatively computationally inexpensive, it can serve as a pre-analysis, making more involved data worth analyses more efficient. The work presented here is a first examination of applications of the uMLR-RDM and the predictive uMLR approaches. There are many other areas to investigate, including: understanding the effects of measurement uncertainty on uMLR results and examining the efficiency of uMLR for more complex conditions and for larger candidate observations sets. Additionally, for predictive analysis, uMLR with RDM is currently able to identify potential best measurement sets; but it is unclear how the correlation coefficients for this set should be determined for any specific field conditions. Finally, it could be useful to examine uMLR for parameter estimation under conditions that are known to be challenging for inverse analyses to see if uMLR could improve the value of the data that is collected.

Supporting Information

Contents of this file

Tables S1 to S4

Introduction

The supporting information included in this document provides details of the multiple linear regression model. Each table contains regression coefficients and error results for the four primary analyses. The four analyses included are: 1) individual station best sets, using downsampled data (section 3.1; Table S1), RDM best sets (implemented at each station), using downsampled data (section 3.1; Table S2), individual soil best sets, using simulated data (section 3.2; Table S3), and RDM best sets (from HYDRUS) implemented at each station, using downsampled data (section 3.3; Table S4).

The coefficients (a_0-a_3) are used in Equation 2 to calculate the time-varying total water storage (predicted). The error values are calculated as length of water [mm] using the predicted and measured storage values. The captions of each table discuss the conditions of the uMLR analysis.

Table 1: Measured Data: Individual Station Best Set

		uMLR Weights				Observation Depth (cm)			Storage Error (mm)		
Condition	**Station**	**a_0**	**a_1**	**a_2**	**a_3**	**Obs1**	**Obs2**	**Obs3**	**Max Error**	**Avg Error**	**RMSE**
Wet	1	5.7	523.2	589.3	661.9	30	90	150	14.0	2.8	3.9
	2	2.6	533.8	785.9	550.8	30	90	165	15.1	3.9	5.5
	3	12.7	461.3	527.5	483.0	30	75	135	16.8	5.9	7.1
	4	9.7	470.9	689.1	572.5	30	90	150	17.9	5.3	7.0
	5	8.1	713.5	577.9	474.4	45	105	165	16.9	3.9	5.4
	6	9.0	692.5	583.2	459.9	45	105	165	17.9	3.8	5.3

Typical	1	7.9	273.7	504.0	945.8	15	60	120	6.9	1.7	2.4
	2	-6.3	279.5	517.9	1159.4	15	60	135	5.3	1.8	2.4
	3	8.3	208.4	452.2	943.9	15	45	120	6.4	2.0	2.7
	4	10.1	219.7	457.7	974.3	15	45	120	6.1	1.7	2.2
	5	32.4	231.3	398.7	675.9	15	45	90	6.3	2.1	2.5
	6	0.1	302.3	541.6	1031.8	15	60	135	8.4	1.7	2.7
	7	35.0	224.1	384.7	686.1	15	45	90	3.8	1.6	1.9
	8	0.6	261.9	339.8	1228.0	15	45	150	4.1	1.7	1.9
	9	30.9	218.2	417.8	687.5	15	45	90	4.4	1.3	1.6
	10	53.5	177.7	332.3	359.0	0	30	60	6.1	2.5	2.9
	11	45.5	159.1	401.5	495.3	0	30	75	8.1	2.5	3.1
	12	28.0	228.6	445.5	656.4	15	45	105	5.0	1.9	2.4

Table S1. For each station the best three observations were found using uMLR for both the wet and typical precipitation conditions (section 3.1). The best observation depths [cm] are given in columns 7-9 (Obs1-Obs3). The water content data at these depths (e.g. Figure 3a,b) and the coefficients in columns 3-6 (a_0-a_3) are used in Equation 2 to calculate the predicted storage value. The predicted storage and the measured storage values (e.g. Figure 3c) are used to calculate the error values in columns 10-12 (absolute maximum error, absolute average error, and RMSE).

Table 2: Measured Data: RDM Best Set	
uMLR Weights	**Error (mm)**

Condition	Station #	a_0	a_1	a_2	a_3	MaxError	AvgError	RMSE
Wet: 30cm 75cm 135cm	1	-6.29	433.86	520.63	882.03	26.19	5.07	7.01
	2	15.44	429.86	521.80	709.43	17.59	5.79	7.22
	3	12.72	461.29	527.53	482.98	16.83	5.86	7.10
	4	14.19	400.17	537.16	676.14	19.61	6.20	7.88
	5	13.45	429.41	492.03	754.92	22.34	5.63	6.99
	6	9.47	394.47	525.81	776.79	17.44	6.22	7.75
Typical: 15cm 60cm 165cm	1	-0.75	303.66	508.04	1172.45	7.48	2.01	2.66
	2	0.64	282.43	522.06	1100.00	7.52	1.78	2.44
	3	13.02	291.33	522.01	961.82	8.02	2.85	3.61
	4	3.98	272.00	508.44	1079.68	6.20	1.90	2.62
	5	-1.31	309.29	476.63	1211.28	7.41	2.48	3.05
	6	13.04	303.40	536.14	880.31	7.04	2.50	3.23
	7	9.36	277.43	578.19	878.65	7.10	2.47	3.23
	8	-3.81	289.69	461.04	1206.65	5.33	1.53	2.13
	9	8.18	289.83	491.93	948.32	7.34	2.15	2.95
	10	-2.64	270.73	489.59	1209.88	7.26	2.07	2.80
	11	4.62	294.15	524.14	1028.06	9.41	2.34	3.35
	12	30.88	293.56	547.12	475.55	7.27	2.76	3.52

Table S2. For each station, the best three observations across all stations (RDM-uMLR) were found for both wet and typical precipitation conditions (section 3.1). The condition and best observation depths [cm] are given in column 1. The water content data (eg. Figure 3a,b) at these depths and the coefficients in columns 3-6 (a_0-a_3) are used in Equation 2 to calculate the predicted storage value. The predicted storage and the measured storage values (e.g. Figure 3c) are used to calculate the error values in columns 7-9 (absolute maximum error, absolute average error, and RMSE).

Table 3: HYDRUS: RDM Best Set - HYDRUS Storage		
	uMLR Weights	**Error (mm)**

Condition	Soil Type	a_0	a_1	a_2	a_3	MaxError	AvgError	RMSE
Wet: 45cm 90cm 150cm	Silt	53.53	492.72	565.25	508.51	21.93	4.10	6.12
	Loam	35.64	495.18	557.33	548.10	23.33	4.20	6.58
	Given	39.28	481.98	518.06	567.16	18.83	5.62	7.34
	Best	9.63	548.61	492.14	700.87	11.22	2.71	3.63
Typical: 30cm 75cm 135cm	Silt	8.78	444.16	426.09	853.08	9.09	1.68	2.77
	Loam	8.47	433.27	428.44	849.06	11.04	2.19	2.94
	Given	2.39	422.52	420.88	921.98	13.27	2.99	3.83
	Best	27.62	415.30	559.68	518.11	7.59	1.35	2.13
	Sand	27.62	415.30	559.68	518.11	6.55	1.57	2.21

Table S3. For each soil type, the best three observations across all stations (RDM-uMLR) were found for both wet and typical precipitation conditions (section 3.2). The condition and best observation depths [cm] are given in column 1. The water content data (Figure 4,5) at these depths and the coefficients in columns 3-6 (a_0-a_3) are used in Equation 2 to calculate the predicted storage value. The predicted storage and simulated storage values (Figure 6) are used to calculate the error values in columns 7-9 (absolute maximum error, absolute average error, and RMSE).

Table 4: HYDRUS: RDM Best Set - Measured Storage

uMLR Weights						Error (mm)		
Condition	**Station #**	a_0	a_1	a_2	a_3	**MaxError**	**AvgError**	**RMSE**
Wet: 45cm 90cm 150cm	1	6.62	634.43	430.60	676.06	16.46	3.71	4.67
	2	11.18	663.01	459.19	596.46	18.55	5.19	6.61
	3	4.15	659.57	479.15	574.14	23.11	7.30	8.70
	4	12.31	615.82	512.41	552.08	18.73	5.82	7.24
	5	11.51	635.50	481.31	590.96	14.78	5.01	6.22
	6	7.10	614.66	534.55	625.57	18.75	5.00	6.25
	1	0.44	448.00	471.70	884.51	6.53	1.83	2.35

	2	6.97	470.94	427.86	805.18	12.06	2.60	3.50
	3	28.33	426.74	567.19	314.00	8.11	2.51	3.32
	4	19.78	446.08	540.56	499.63	10.57	2.99	4.01
	5	8.19	472.45	468.34	750.57	10.10	2.18	3.05
Typical: 30cm 75cm 135cm	6	2.83	452.04	420.63	895.00	6.63	1.97	2.66
	7	9.35	422.41	462.46	727.04	6.32	2.01	2.71
	8	-1.44	392.68	475.60	954.65	30.23	4.33	6.94
	9	14.98	440.38	523.19	580.02	10.88	2.52	3.56
	10	8.62	397.08	508.68	721.14	4.92	1.59	2.06
	11	0.11	464.11	413.51	944.09	8.59	2.40	3.03
	12	29.35	444.18	648.78	199.49	10.66	2.53	3.39

Table S4. For each soil type, the best three observations across all stations (RDM-uMLR) were found for both wet and typical precipitation conditions (section 3.2). The condition and best observation depths [cm] are given in column 1. Next, the uMLR results for measured data (at the HYDRUS recommended depths) were used to find the coefficients in columns 3-6 (a0-a3) (section 3.3). The measured water content data (e.g. Figure 3a,b) at these depths and the coefficients are used to calculate (Equation 2) the predicted storage value. The predicted storage and measured storage values (e.g. Figure 3c) are used to calculate the error values in columns 7-9 (absolute maximum error, absolute average error, and RMSE).

References

Anderson, M. P. (2005). Heat as a ground water tracer. *Groundwater*, *43*(6), 951-968.

Bakker, M., Post, V., Langevin, C. D., Hughes, J. D., White, J. T., Starn, J. J. and Fienen, M. N., 2016, Scripting MODFLOW model development using Python and FloPy: Groundwater, doi:10.1111/gwat.12413

Bredehoeft, J. D., & Papaopulos, I. S. (1965). Rates of vertical groundwater movement estimated from the earth's thermal profile. *Water Resources Research*, *1*(2), 325-328.

Çamdevýren, H., Demýr, N., Kanik, A., & Keskýn, S. (2005). Use of principal component scores in multiple linear regression models for prediction of Chlorophyll-a in reservoirs. *Ecological Modelling*, *181*(4), 581-589.

Chung, S. O., & Horton, R. (1987). Soil heat and water flow with a partial surface mulch. *Water Resources Research*, *23*(12), 2175-2186.

Clutter, M., & Ferré, T. (2018). Examining the Potentials and Limitations of Using Temperature Tracing to Infer Water Flux through Unsaturated Soils. *Vadose Zone Journal*, *17*(1), doi:10.2136/vzj2017.10.0181.

Clutter, M.J. & Ferre, P.A. (in press). Designing robust, cost-effective field measurement sets using universal multiple linear regression. *Soil Science Society of America Journal*.

Clutter, M.J. & Ferre, P.A. (in submission). Robust predictive design of field measurements for evapotranspiration barriers using universal multiple linear regression. *Water Resources Research*.

Constantz, J., Tyler, S. W., & Kwicklis, E. (2003). Temperature-profile methods for estimating percolation rates in arid environments. *Vadose Zone Journal*, *2*(1), 12-24.

Dickinson, J. E., Ferré, T. P. A., Bakker, M., & Crompton, B. (2014). A screening tool for delineating subregions of steady recharge within groundwater models. *Vadose Zone Journal*, *13*(6), vzj2013.10.0184.

DOE-RL. 2016. *Prototype Hanford Barrier 1994 to 2015,* DOE/RL-2016-37, Rev. 0, U.S. Department of Energy Richland Operations Office. Richland, Washington. Available at https://www.hanford.gov/c.cfm/sgrp/DOE-RL-2016-37/DOE-RL-2016-37_R0.pdf and https://www.hanford.gov/c.cfm/sgrp/DOE-RL-2016-37/Appendices.pdf (Accessed on March 19, 2019).

Doherty, J. (2016). PEST Model-Independent Parameter Estimation User Manual Part I: PEST, SENSAN and Global Optimisers. *Watermark Numerical Computing*, *6*.

Draper, N. R., & Smith, H. (2014). *Applied regression analysis* (Vol. 326). John Wiley & Sons.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans* (Vol. 38). Siam.

Fayer, M. J., & Keller, J. M. (2007). *Recharge Data Package for Hanford Single-Shell Tank Waste Management Areas* (No. PNNL-16688). Pacific Northwest National Lab (PNNL), Richland, WA (United States).

Feddes, R. A., Kowalik, P., Kolinska-Malinka, K., & Zaradny, H. (1976). Simulation of field water uptake by plants using a soil water dependent root extraction function. *Journal of Hydrology*, *31*(1-2), 13-26.

Fernández-Gálvez, J., Simmonds, L. P., & Barahona, E. (2006). Estimating detailed soil water profile records from point measurements. *European Journal of Soil Science*, *57*(5), 708-718.

Gardner, W., & Kirkham, D. (1952). Determination of soil moisture by neutron scattering. *Soil Science*, *73*(5), 391-402.

Gee, G. W., Ward, A. L., Gilmore, B. G., Ligotke, S. O., Link, S.O. (1995). *Hanford prototype-barrier status report FY 1995* (No. PNNL-10872). Pacific Northwest National Lab (PNNL)., Richland, WA (United States).

Gee, G. W., Ward, A. L., Gilmore, B. G., Link, S. O., Dennis, G. W., & ONeil, T. K. (1996). *Hanford prototype-barrier status report FY 1996* (No. PNNL-11367). Pacific Northwest National Lab (PNNL)., Richland, WA (United States).

Groves, D. G., Davis, M., Wilkinson, R., & Lempert, R. (2008). Planning for climate change in the Inland Empire: Southern California. *Water Resources IMPACT*, *10*(4), 14-17.

Gupta, H. V., Sorooshian, S., & Yapo, P. O. (1998). Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Research*, *34(4),* 751-763.

Gupta, H. V., Wagener, T., & Liu, Y. (2008). Reconciling theory with observations: elements of a diagnostic approach to model evaluation. *Hydrological Processes, 22(18),* 3802-3813.

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., & Ye, M. (2012). Towards a comprehensive assessment of model structural adequacy. *Water Resources Research, 48(8),* doi:10.1029/2011WR011044.

Harbaugh, A.W., 2005, MODFLOW-2005, the U.S. Geological Survey modular ground-water model -- the Ground-Water Flow Process: U.S. Geological Survey Techniques and Methods 6-A16.
This report describes the theory and input instructions at the time of the initial MODFLOW-2005 v1.00 release.

Haykin, S. (1994). *Neural Networks: A Comprehensive Foundation*. New Jersey: Prentice Hall PTR.

Hermans, T., Nguyen, F., & Caers, J. (2015). Uncertainty in training image-based inversion of hydraulic head data constrained to ERT data: Workflow and case study. *Water Resources Research*, *51*(7), 5332-5352.

Hine, D., & Hall, J. W. (2010). Information gap analysis of flood model uncertainties and regional frequency analysis. *Water Resources Research*, *46*(1), doi:10.1029/2008WR007620.

Hoitink, D. J., Burk, K. W., Ramsdell, J. V., & Shaw, W. J. (2005). *Hanford Site Climatological Summary 2004 with Historical Data* (No. PNNL-15160). Pacific Northwest National Lab (PNNL)., Richland, WA (United States).

Høybye, J. A. (1998). Model error propagation and data collection design. An application in water quality modelling. *Water, Air, and Soil Pollution*, *103*(1-4), 101-119.

Hubbart, J., Link, T., Campbell, C., & Cobos, D. (2005). Evaluation of a low-cost temperature measurement system for environmental applications. *Hydrological Processes*, *19*(7), 1517-1523.

James, B. R., & Gorelick, S. M. (1994). When enough is enough: The worth of monitoring data in aquifer remediation design. *Water Resources Research*, *30*(12), 3499-3513.

Jolliffe, I. (2011). Principal component analysis, *International encyclopedia of statistical science* (pp. 1094-1096). Berlin/Heidelberg: Springer.

Ju, L. , Zhang, J. , Wu, L. and Zeng, L. (2018), Bayesian Monitoring Design for Streambed Heat Tracing: Numerical Simulation and Sandbox Experiments. *Groundwater*. doi:10.1111/gwat.12823.

Keery, J., Binley, A., Crook, N., & Smith, J. W. (2007). Temporal and spatial variability of groundwater–surface water fluxes: development and application of an analytical method using temperature time series. *Journal of Hydrology*, *336*(1-2), 1-16.

Kikuchi, C. (2017). Toward Increased Use of Data Worth Analyses in Groundwater Studies. *Groundwater*, *55*(5), 670-673.

Kirkham, D., & Powers, W. L. (1972). *Advanced soil physics*. New York: Wiley.

Kurc, S. A., & Small, E. E. (2004). Dynamics of evapotranspiration in semiarid grassland and shrubland ecosystems during the summer monsoon season, central New Mexico. *Water Resources Research*, *40(9)*, doi:10.1029/2004WR003068.

Lempert, R. J., & Groves, D. G. (2010). Identifying and evaluating robust adaptive policy responses to climate change for water management agencies in the American west. *Technological Forecasting and Social Change*, *77*(6), 960-974.

Lempert, R. J., Popper, S. W., Groves, D. G., Kalra, N., Fischbach, J. R., Bankes, S. C., Bryant, B. P., Collins, M. T., Keller, K., Hackbarth, A., Dixon, L., LaTourrette, T., Reville, R. T., Hall, J. W., Mijere, C., & McInerney, D. J. (2013). *Making Good Decisions Without Predictions: Robust Decision Making for Planning Under Deep Uncertainty.* Retrieved from https://www.rand.org/pubs/research_briefs/RB9701.html.

Lin, Y., Le, E. B., O'Malley, D., Vesselinov, V. V., & Bui-Thanh, T. (2017). Large-scale inverse model analyses employing fast randomized data reduction. *Water Resources Research, 53(8),* 6784-6801.

Liu, X., Lee, J., Kitanidis, P. K., Parker, J., & Kim, U. (2012). Value of information as a context-specific measure of uncertainty in groundwater remediation. *Water Resources Management, 26(6),* 1513-1535.

Massmann, J. W. (1989). Applying groundwater flow models in vapor extraction system design. *Journal of Environmental Engineering*, *115*(1), 129-149.

Minsker, B. (2003). Long-Term Groundwater Monitoring–The State of the Art. In *American Society of Civil Engineers*.

Mogheir, Y., Singh, V. P., & de Lima, J. L. M. P. (2003). Redesigning the Gaza Strip groundwater quality monitoring network using entropy, *Ground water pollution* (pp. 315-331). New Delhi: Allied.

Norgaard, T., Moldrup, P., Ferré, T. P. A., Katuwal, S., Olsen, P., & De Jonge, L. W. (2014). Field-scale variation in colloid dispersibility and transport: Multiple linear regressions to soil physico-chemical and structural properties. *Journal of Environmental Quality*, *43(5)*, 1764-1778.

Popper, S. W., Berrebi, C., Griffin, J., Light, T., Min, E., & Crane, K. (2009). *Natural gas and Israel's energy future: Near-term decisions from a strategic perspective* (Vol. 927). Rand Corporation.

Reed, P., Minsker, B., & Valocchi, A. J. (2000a). Cost-effective long-term groundwater monitoring design using a genetic algorithm and global mass interpolation. *Water Resources Research*, *36*(12), 3731-3741.

Reed, P., Minsker, B., & Goldberg, D. E. (2000b). Designing a competent simple genetic algorithm for search and optimization. *Water Resources Research*, *36*(12), 3757-3761.

Rickard, W. H., and B. E. Vaughan (1988), Plant community characteristics and responses, in Shrub-Steppe: Balance and Change in a Semi-Arid Terrestrial Ecosystem, Developments in Agricultural and Managed-Forest Ecology, edited by W. H. Rickard, B. E. Vaughan, and L. E. Rogers,pp. 109–179, Elsevier, Amsterdam, Netherlands.

Rousseau, J. P., Kwicklis, E. M., & Gillies, D. C. (1999). *Hydrogeology of the unsaturated zone, North Ramp area of the exploratory studies facility, Yucca Mountain, Nevada* (No. 98-4050). US Geological Survey.

Scanlon, B. R., Levitt, D. G., Reedy, R. C., Keese, K. E., & Sully, M. J. (2005). Ecological controls on water-cycle response to climate variability in deserts. *Proceedings of the National academy of Sciences*, *102*(17), 6033-6038.

Schaap, M. G., Leij, F. J., & Van Genuchten, M. T. (2001). Rosetta: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. *Journal of Hydrology*, *251*(3-4), 163-176.

Sela, S., Svoray, T., & Assouline, S. (2015). The effect of soil surface sealing on vegetation water uptake along a dry climatic gradient. *Water Resources Research*, *51*(9), 7452-7466.

Shan, C., & Bodvarsson, G. (2004). An analytical solution for estimating percolation rate by fitting temperature profiles in the vadose zone. *Journal of Contaminant Hydrology*, *68*(1-2), 83-95.

Šimunek, J., Van Genuchten, M. T., & Sejna, M. (2005). The HYDRUS-1D software package for simulating the one-dimensional movement of water, heat, and multiple solutes in variably-saturated media. *University of California-Riverside Research Reports*, *3*, 1-240.

Stallman, R. W. (1965). Steady one-dimensional fluid flow in a semi-infinite porous medium with sinusoidal surface temperature. *Journal of Geophysical Research*, *70*(12), 2821-2827.

Strebelle, S. (2002). Conditional simulation of complex geological structures using multiple-point statistics. *Mathematical Geology*, *34*(1), 1-21.

Sun, Y., Hou, Z., Huang, M., Tian, F., & Ruby Leung, L. (2013). Inverse modeling of hydrologic parameters using surface flux and runoff observations in the Community Land Model. *Hydrology and Earth System Sciences*, *17*(12), 4995-5011.

Tiedeman, C. R., Hill, M. C., D'Agnese, F. A., & Faunt, C. C. (2003). Methods for using groundwater model predictions to guide hydrogeologic data collection, with application to the Death Valley regional groundwater flow system. *Water Resources Research, 39(1),* doi:10.1029/2001WR001255.

Tonkin, M.J., Tiedeman C. R., Ely D.M., and Hill M.C., 2007, OPR-PPR, a Computer Program for Assessing Data Importance to Model Predictions Using Linear Statistics: Reston Virginia, U.S. Geological Survey Techniques and Methods Report TM-6E2, 115 pages. http://pubs.usgs.gov/tm/2007/tm6e2/.

Topp, G. C., Davis, J. L., & Annan, A. P. (1980). Electromagnetic determination of soil water content: Measurements in coaxial transmission lines. *Water Resources Research*, 16(3), 574-582.

USDOE. 1987. Final environmental impact statement: Disposal of Hanford defense high-level, transuranic and tank wastes. USDOE, Richland, WA.

van Genuchten, M. T. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils 1. *Soil Science Society of America Journal*, *44*(5), 892-898.

Vereecken, H., Huisman, J. A., Bogena, H., Vanderborght, J., Vrugt, J. A., & Hopmans, J. W. (2008). On the value of soil moisture measurements in vadose zone hydrology: A review. *Water Resources Research*, *44*(4), doi:10.1029/2008WR006829.

Vilhelmsen, T. N., & Ferré, T. P. (2018). Extending Data Worth Analyses to Select Multiple Observations Targeting Multiple Forecasts. *Groundwater*, *56*(3), 399-412.

Wagner, B. J. (1995). Sampling design methods for groundwater modeling under uncertainty. *Water Resources Research*, *31*(10), 2581-2591.

Walker, W. W. (1982). A Sensitivity and Error Analysis Framework for Lake Eutrophication Modeling 1. *JAWRA Journal of the American Water Resources Association*, *18*(1), 53-60.

Walvoord, M. A., Scanlon, B. R., Iogan, J. F., & Phillips, F. M. (2004). Hydrologic processes in deep vadose zones in interdrainage arid environments. *Groundwater Recharge in a Desert Environment: The Southwestern United States.* Washington DC: American Geophysical Union.

Wilcox, B. P., Seyfried, M. S., Breshears, D. D., Stewart, B., & Howell, T. (2003). The water balance on rangelands. *Encyclopedia of water science*, 791-794.

Zhang, Z. F. (2015). Field soil water retention of the Prototype Hanford Barrier and its variability with space and time. *Vadose Zone Journal*, *14*(8).

Zhang, Z. F. (2016). Evaluating the long-term hydrology of an evapotranspiration-capillary barrier with a 1000 year design life. *Water Resources Research*, *52*(6), 4883-4904.